# IBERGRID 2019 - Delivering Innovative Computing and Data services to Researchers

Monday, September 23, 2019 - Thursday, September 26, 2019

# Book of Abstracts

# Contents

**e-Infrastructure Plenaries** / **2**

# IBERGRID

**e-Infrastructure Plenaries** / **3**

# EOSC-synergy: Expanding the capacity and capabilities of EOSC at the National levels

**e-Infrastructure Plenaries** / **4**

# The Research Data Alliance: a (research) data window to the world

**IBERGRID Contributions** / **5**

# Using Big Data for Anomaly Detection

**Author:** Javier Cacheiro[1]

[1] *CESGA*

**Corresponding Author:** jlopez@cesga.es

During the last years, Big Data technologies, and in particular Hadoop and HBase, have enabled us to expand enormously the information that we collect and store from all our servers and infrastructures. We no longer need to discard old data using round-robin-databases or restrict the number of active nagios-style checks.

Now we can take full advantage of a metric collection infrastructure that allows to perform nagios-style checks directly against the metrics database instead of directly accessing the servers.

The information currently includes tens of thousands of time-series that are stored in HBase as well as a large collection of logs stored in HDFS.

The next challenge, is to analyze this data to detect anomalous behaviour, usually this is done by the operators looking at different operational dashboards, however data anomaly set has become too large and diverse for manual interpretation.

To take advantage of these metrics, we started evaluating generic anomaly detection techniques, applying them directly to our time-series and log data. The main problem we encountered when evaluating these generic solutions is that they produce a large number of false positives that greatly reduce their usefulness. It is not practical to have a system that produces so many alerts than it is impossible for operators to investigate all of them.

So three years ago, we started developing our own custom algorithms to detect anomalies. We will show how this approach enabled us not only to have a better understanding of our systems but also to obtain accurate results for different use cases ranging from SSH attack detection to CPU malfunctioning detection.

**IBERGRID Contributions / 6**

## Serverless Computing for Data-Processing Across Public and Federated Clouds

**Authors:** Sebastián Risco[1]; Alfonso Pérez González[2]; Miguel Caballer[1]; Germán Moltó[1]

[1] *Universitat Politècnica de València*

[2] *UPV - GRyCAP*

**Corresponding Authors:** gmolto@dsic.upv.es, serisgal@i3m.upv.es, micafer1@upv.es, alpegon3@upv.es

Serverless computing is evolving from the initial Functions as a Service (FaaS) approach to also embrace the execution of containerised applications without the user managing the underlying computing infrastructure. Indeed, the main public cloud providers such as Amazon Web Services or Google Cloud have already started to offer services in this regard. This is the case of AWS Fargate or Google Cloud Run, mainly aimed at the deployment of microservices-based architectures. However, scientific computing can also benefit from the elastic automated management of computational infrastructure for data processing. To this aim, we developed SCAR, an open-source framework to run containers out of Docker images on AWS Lambda which defines a file-processing computing model that is triggered in response to certain events (such as file upload or a REST API invocation). This model was extended for on-premises environments through OSCAR, an open-source platform which enables the users to deploy their file-processing container-based serverless applications on a dynamically provisioned elastic Kubernetes cluster that can be deployed in multi-Clouds, integrated with the EGI Federated Cloud and the EGI Data Hub, based on Onedata.

In this work we focus on integrating a federated storage for data persistence, in particular the EGI Data Hub, with the ability to dynamically provision computational resources from a public Cloud provider to perform the data processing in response to file uploads. To this aim, we developed OneTrigger, a tool to trigger events from Onedata, that can be run as a serverless function in AWS Lambda in order to use SCAR's functionality to perform the execution of jobs in AWS Lambda, supporting thousands of concurrent executions. Longer executions, as well as those requiring specialised computing hardware, such as GPUs, are delegated to AWS Batch, a service which enables the unattended and elastic execution of batch computing workloads on the public Cloud. This allows to create hybrid data-processing serverless applications across public and federated Clouds. We demonstrate the feasibility of this approach by introducing a use case in video processing that can leverage GPU-based computing in the public Cloud to dramatically accelerate object recognition, while data persistence is still supported by the federated Cloud.

**IBERGRID Contributions / 7**

## Computational challenges related to IFMIF and DONES facilities.

**Author:** Ortiz Christophe J. [1]

[1] *CIEMAT*

Following ITER, DEMO reactor is expected to demonstrate the feasibility of safe, environmentally friendly and economically viable fusion power generation. During operation of DEMO, the materials will be exposed to a particular hostile environment as a consequence of the energetic neutrons created by fusion reactions in the plasma. The level of damage expected in fusion conditions is such that the performance of materials and components under these extreme irradiation conditions is unknown. One of the central objectives of the fusion materials program is to identify innovative materials development routes, using scientific understanding and knowledge of how materials properties evolve and change in the operating environment of a fusion power plant.
In this respect, IFMIF is considered as one of the main pillars in the international fusion program. Its double deuteron beam 125 mA each will produce enough rate of damage behind

the lithium target to make available in a few years information on materials damage at DEMO relevant doses. On the other hand, DONES (DEMO Oriented Neutron Source) has been conceived as a simplified IFMIF-like plant to provide in a reduced time scale and with a reduced budget – both compared to IFMIF- the basic information on materials damage. Although both facilities are designed to provide experimental data on how the material properties change under energetic neutron irradiation, the design of experiments to be carried out to test materials implies various computational challenges. During our talk we shall review the different computational fields associated to IFMIF and DONES facilities, such as beam dynamics, neutronic transport, calculation of collision cascades and the simulation of the microstructure evolution in the irradiated materials.

**IBERGRID Contributions / 8**

## Cosmology @EOSC: the HPC Universe in the Cloud

**Author:** Francisco Prada[1]

[1] *IAA-CSIC*

The new generation of upcoming galaxy surveys will measure the effect of dark energy on the expansion history of the universe. They will obtain in the next decade dozens of millions of galaxy data, constructing a 3-D map spanning the nearby universe to 10 billion light-years, and will provide an accurate determination of the distance-redshift relation. Extracting cosmological information on the nature of the dark matter-energy components of the universe and unveiling new physics from these experiments requires to run highly demanding HPC cosmological simulations with extraordinary high numerical resolution for a huge volume. The new Uchuu N-body simulation (Universe in Japanese) meets all these requirements. The simulation is finished and we have been able to obtain all necessary results from the extensive analysis of the Uchuu raw particle data (more than 6 petabytes!) to generate galaxy formation semi-analytical models and gravitational lensing maps in order to produce high-fidelity galaxy mocks that are close to what will be observed by those large surveys. The final Uchuu products can only be disseminated to the public in a Cloud Computing Platform. I will give an overview about the science context, impact of the project, description of the products, analysis tools, dissemination plan and cloud support needs.

**IBERGRID Contributions / 9**

## Using HPC to enable coastal waters observatories

**Authors:** Marta Rodrigues[1]; Anabela Oliveira[1]; Ricardo Martins[1]; João Rogeiro[1]; Daniela Santos[1]; André Fortunato[1]; Alberto Azevedo[1]

[1] *Laboratório Nacional de Engenharia Civil*

**Corresponding Authors:** rjmartins@lnec.pt, mfrodrigues@lnec.pt, aoliveira@lnec.pt, dasantos@lnec.pt, aazevedo@lnec.pt, afortunato@lnec.pt, jrogeiro@lnec.pt

Coastal systems are among the most productive ecosystems in the world, providing multiple resources and guaranteeing the resilience of the coastal communities. Climate change (e.g., sea level rise) represents a major threat to the world's coastal systems, via potential increases in salinity, acceleration in the nutrients cycling and disruption of aquatic ecosystems. Also, recent and predicted increases of nutrients loads to coastal systems may exacerbate these impacts.
Coastal waters observatories can support both the daily and the long-term management of coastal ecosystems, allowing the continuous surveillance of coastal zones and the establishment of adaptation measures. In the project UBEST the concept of coastal waters observatories is extended and demonstrated in two Portuguese coastal systems, the Tagus estuary and the Ria Formosa, to improve the global understanding of the biogeochemical buffering capacity of coastal ecosystems and their susceptibility to future scenarios of anthropogenic inputs and climate change.

The observatories developed in UBEST include several layers of information that integrate historical and real-time observations, forecasts, scenarios analysis and indicators in a comprehensive web-portal. The integration of all these layers provides information that covers different temporal scales, presented with different levels of complexity, enabling the end-users with more robust tools to support decision-making. However, the extension of the coastal waters observatories to integrate more layers of information brings several challenges, among them the requirement of more computational resources. In this context, High Performance Computing (HPC) is a powerful resource to enable the next generation of coastal waters observatories.
HPC, such as grid clusters, parallel computing or cloud computing, is used by the coastal modeling community to solve complex, very demanding problems. In UBEST, HPC is used at two levels: i) for high-resolution forecasts and scenarios simulations of the circulation and water quality dynamics in the two coastal systems, and ii) to provide computational power to process data and model results through predefined or user requests at the web-portal.
The simulations in the Tagus estuary and the Ria Formosa are performed with SCHISM, a parallelized model that uses the MPI (Message Passing Interface) paradigm. Daily forecasts of water levels and 3D currents, salinity, temperature and biogeochemical variables are deployed with the WIFF – Water Information and Forecasting Framework and the OPENCoastS service. The scenarios analysis provides long-term information of the biogeochemical buffering capacity of each system under present conditions and for scenarios of climate change (e.g. sea level rise) and anthropogenic pressures (e.g. wastewater discharges). The use of HPC allows both the timely production of daily forecasts and the generation of long-term simulations for the scenarios.
The UBEST web-portal, developed using Django, allows the access to all the data and model results through four dashboards: Data, Forecasts, Scenarios and Indicators dashboards. Several services and products are made available to the users, such as statistics of historical data, data on virtual sensors, and physical and water quality indicators.
The implementation of HPC in the UBEST water observatories was achieved using the INCD – the Portuguese National Infrastructure for Distributed Computing.

**IBERGRID Contributions / 10**

# hybrid batch system deployment with AWS spot instances

**Authors:** Carles Acosta[1]; Gonzalo Merino[2]; Jordi Casals[2]; Vanessa Acín[1]

[1] *IFAE*

[2] *CIEMAT*

**Corresponding Authors:** vacin@pic.es, jcasals@pic.es, cacosta@pic.es, merino@pic.es

Our institution, Port d'Informació Científica (PIC), is an innovative centre for supporting research and provides support to scientific groups working in projects which require large amount of computing resources for the analysis of massive sets of distributed data. PIC is the Spanish Tier-1 center for the Large Hadron Collider, the main (Tier-0) data center for the MAGIC telescopes and the PAU dark energy survey, and is one of the Science Data Centers of ESA's Euclid mission.

At PIC we have piloted a hybrid cloud computing platform totally integrated in our batch computing service and transparent to the final users. We doubled our computing capacity using AWS spot instances for 72 hours in order to test how we can increase our peak computing needs at an affordable price.

To test this hybrid batch system infrastructure we have used the HTCondor condor_annex tool, which makes the process of extending a local pool with cloud resources easy, fast and if the user needs it, with an expiration date. In order to get to the production ready system, everything was tested in three steps: small batch of on-demand instances in a test environment, small batch of on-demand and spot-instances in a production environment and big batch of spot instances in a production environment.

Initially the jobs were sent to our test environment to then be moved to production after checking that the jobs were running correctly, both of them using on-demand instances. The test continued by launching spot-instances in a seamless hybrid infrastructure where the cloud worker nodes were

added to the local computing pool and have jobs running in minutes. Accounting and monitoring of the cloud resources has been totally integrated with the local system.

Amazon Web Services Spot Instances offers the possibility to instantiate machines at a fraction of the on-demand price due to low demand of specific instance types at specific times. When a lot of instances are launched and the conditions to keep them running change, some or all of them can be stopped at any moment. This suits very well use cases like the one tested here.

There were some other elements needed to configure the system, such as a custom worker node image created and stored in a specific region in AWS or a HTCondor Connection Broker (CCB) to enable communication between the AWS nodes and the local system, apart from the changes in the HTCondor configuration to accept the new servers as own.

**IBERGRID Contributions / 11**

## Experience with the GÉANT Cloud IaaS Framework Agreement

**Authors:** André Vieira[1]; Mário David[2]; João Martins[2]

[1] *INCD*

[2] *LIP / INCD*

**Corresponding Authors:** andrev@lip.pt, david@lip.pt, martinsj@lip.pt

GÉANT has carried out a European wide Framework Procurement for an Infrastructure as a Service (IaaS) cloud portfolio for the European research and education sector. The result was a multi-supplier framework whereby a number of IaaS cloud vendors were awarded framework contracts. Under this framework academic and research organizations from European Union countries can directly contract cloud IaaS services from these vendors through a simplified purchase procedure, while respecting the national public procurement laws of member countries. Within each country the national research network (NREN) is involved in the promotion of the agreement. In order to validate the framework agreement several organizations have been piloting services contracted under the agreement.

In this communication we describe the piloting activities conducted by the Portuguese Distributed Computing Infrastructure (INCD) aimed at validating and exploiting the capabilities of the GÉANT cloud framework agreement in Portugal. The following aspects will be addressed: legal compliance of the framework to the national public purchase laws; contractual process from the selection of a vendor to the actual service usage; comparison of service offerings and related conditions both against the vendor commercial conditions and against in-house service provisioning; assessment of the billing and payment processes; evaluation of the actual service delivery and support.

**IBERGRID Contributions / 12**

## Rootless containers with udocker

**Authors:** Jorge Gomes[1]; Mário David[2]; João Pina[1]; João Martinsj[1]

[1] *LIP / INCD*

[2] *LIP*

**Corresponding Authors:** jorge@lip.pt, david@lip.pt, martinsj@lip.pt, jpina@lip.pt

udocker (https://github.com/indigo-dc/udocker) is a tool that addresses the problematic of executing Linux containers in user space, i.e. without installing additional system software, without requiring

administrative privileges and respecting resource usage policies, accounting and process controls. udocker empowers users to execute applications encapsulated in containers easily across a wide range of Linux distributions and systems including computing clusters.

udocker implements a subset of Docker commands aimed at searching, pulling, importing, loading and executing containers. The self installation allows a user to transfer udocker and execute it to pull the required tools and libraries. All required binary tools and libraries are provided with udocker and compilation is not required. udocker is an integration tool that incorporates several execution methods giving the user several options to run their containers according to the host capabilities. Several interchangeable execution modes are available, that exploit different technologies and tools, enabling udocker to run in older and newer Linux distributions. Currently udocker supports four execution modes: system call interception and pathname rewriting via PTRACE, dynamic library call interception and pathname rewriting via shared library preload, Linux unprivileged namespaces via runC, and Singularity when locally available. Each approach has its own advantages and limitations, and therefore an integration tool offers flexibility and freedom of choice to better match the applications to the host characteristics. udocker has more than 500 stars on github and is commonly used to execute HTC, HPC and GPGPU applications across datacenters and infrastructures. udocker was developed by LIP in the context of the INDIGO-DataCloud project and is being further extended in DEEP-Hybrid-DataCloud.

This communication will provide an overview of the udocker capabilities, development status and evolution.

**Welcome & Opening Plenaries** / **13**

## Welcome and Opening

Welcome address by MICIU, FCT, CSIC, EC and CESGA

**Welcome & Opening Plenaries** / **14**

## IBERGRID status presentation

**Corresponding Author:** jorge@lip.pt

**e-Infrastructure Plenaries** / **15**

## The ascent of scientific computing: the EGI role and contribution towards the European Open Science Cloud

This presentation provides an overview of the central role of
distributed data processing to support scientific excellence of
international collaboration in the past decade.
We present the architecture and governance model of EGI, the European infrastructure for exabyte-scale computing, and we demonstrate how open science has been benefiting from the power delivered by the EGI Federation, connecting more than 1,000,000 CPU cores worldwide to realize the largest computing platform for research in the world. The presentation concludes by introducing the technical and organizational challenges that scientific computing will face in the coming decade, and the role that EGI and IBERGRID will play in the context of the European Open Science Cloud initiative of the European Commission.

**Research Communities & EOSC** / **16**

## ESCAPE ESFRI cluster presentation

**Research Communities & EOSC** / **17**

## EXPANDS project presentation

**Research Communities & EOSC** / **18**

## Cosmology @EOSC

**IBERGRID Contributions** / **19**

## Computational challenges related to IFMIF and DONES facilities.

**Innovative Software Services** / **20**

## Serverless: What's in a name for scientific computing?

**Innovative Software Services** / **21**

## IBM-Q - Online Quantum Computing Platform

CSIC and IBM have signed a contract to provide researchers with access to a quantum computer with 20 qubits, via a cloud service. This contract extends the possibilities already existing with the 5-qubit and 16-qubit free tier services, towards devices with greater quantum volume. This talk will offer a superficial overview of quantum computing –operations, implementation, potential–, focusing on how these services are actually offered and used, and how they can integrate in hybrid quantum-classical work pipelines.

**Innovative Software Services** / **22**

## Innovative Software Services

**IBERGRID Contributions / 23**

## Computational challenges related to IFMIF and DONES facilities.

**IBERGRID Contributions / 24**

## Cosmology @EOSC: the HPC Universe in the Cloud

**IBERGRID Contributions / 25**

## RESCCUE RAF app – an IT solution for digital interactive urban resilience assessment

**Authors:** Pedro Lopes[1]; Anabela Oliveira[1]; Cristina Pereira[1]; Rita S. Brito[1]; Maria A. Cardoso[1]; Ricardo Martins[1]; Mário David[2]; Jorge Gomes[2]; João Pina[2]

[1] *Laboratório Nacional de Engenharia Civil*

[2] *LIP*

**Corresponding Authors:** david@lip.pt, jorge@lip.pt, rjmartins@lnec.pt, clpereira@lnec.pt, macardoso@lnec.pt, aoliveira@lnec.pt, plopes@lnec.pt, rsbrito@lnec.pt, jpina@lip.pt

Climate change (CC) adaptation plays an important role in city and services management and resilience building, targeting the mitigation and adaptation to potential hazards in urban areas. Information technologies can play a leading role to promote fast adoption of the most relevant measures towards CC preparedness. In this paper, a web application is presented with the objective of empowering city and services managers with an accessible and reliable tool. The RESCCUE RAF App materializes a detailed CC resilience evaluation methodology with a user-friendly Web interface. It provides an evaluation of city resilience to CC impacts and urban systems vulnerabilities allowing to assess multi-sector dependencies under multiple CC scenarios. This app is integrated as a service of the Portuguese Infrastructures Roadmap, under the Infraestrutura Nacional de Computação Distribuída (INCD) infrastructure initiative that provides to the app the resources for data computation and storage, and assures its scalability to handle multiple user requests as well as database storage growth. The information provided by this app empowers city and urban services managers with an assessment allowing to know where they stand and to identify the resilience gaps, thus supporting decision on the most advantageous investments on the city and services and planning to cope with future challenges. Three case studies are being carried out in different cities (Barcelona, Lisbon and Bristol). The access to the application is made using credentials given upon request, to ensure data confidentiality. Inside the user's area, the user can fill, in an interactive way, detailed information about the selected city, regarding multiple aspects such as financial plan per service, date of last review of the City Master Plan, history of climate hazards in the city or level of dependency between services. This information is then processed and several indicators are calculated on-the-fly. The assessment allows to identify development levels, ranging from the whole city to a more detailed assessment regarding a specific service. Data is stored at INCD's in RESCCUE RAF app database and can be easily analyzed and extracted by the user. These results support the city and services managers in making effective decisions to plan city resilience enhancement. In this paper, a detailed presentation of the architecture and computational choices behind the RESCCUE RAF App and its Web interface and their integration in the INCD infrastructure will be presented. Given its importance, generic nature and flexible structure, the RESCCUE RAF App can be extended to other cities and, in the future, to other urban services or hazards, taking advantage of the INCD e-infrastructure. This complete and in-depth assessment of city resilience to CC challenges at Portuguese, Iberian and European scale is fundamental to plan CC adaptation and strategies implementation, preventing both human and material losses as well as environmental damages.

**IBERGRID Contributions / 26**

# Comparison of Container-based Virtualization Tools for HPC Platforms.

**Authors:** Diana María Naranjo Delgado [1]; Germán Moltó[2]; Jorge Gomes[3]; Mário David[3]; Ignacio Blanquer Espert[2]

[1] *UPV*

[2] *Universitat Politècnica de València*

[3] *LIP*

**Corresponding Authors:** gmolto@dsic.upv.es, dnaranjo@i3m.upv.es, jorge@lip.pt, david@lip.pt, iblanque@dsic.upv.es

Virtualization technologies are a fundamental element in cloud computing. Docker is the most known and used container platform worldwide. It is designed for microservices virtualization and application delivery but its model does not fit well with High-Performance Computing (HPC) platforms. HPC environments are multi-user systems where users should only have access to their own data and computing resources. Misconfigured Docker installations pave the way for privilege escalation, including the ability to access other users' data and, at the same time, gaining control of the cluster and computing resources.

In the world of HPC, the focus of containerised applications is not necessarily on DevOps, but on the ability to minimise HPC node configuration and manage applications' software dependencies through containers. Several open source initiatives have addressed this problem of bringing containers to the HPC space such as Singularity, Shifter, CharlieCloud and uDocker. In this sense, Singularity seems to be the most popular container system for HPC centres, but there are alternatives such as uDocker that support the execution of containers in user space, a key feature in HPC platforms. Therefore, it is important to analyze the benefits and drawbacks of these solutions when they are deployed in real HPC system and applied to scientific production applications.

All these tools, with potentially similar characteristics, bring the benefits of the containers to the HPC world. However, it is important to analyze important metrics in order to determine the advantages of one over another. The fields to analyze include, but are not limited to: interaction with Docker, support for Graphics Processing Unit (GPU), support for low-latency interconnects such as InfiniBand, support for Message Passing Interface (MPI), security and portability, privilege model, integration with Local Resource Management Systems (LRMS), among others. The objective of this communication is to show the behaviour and limitations of different container technologies in the context of HPC systems.

Keywords: virtualization, HPC, uDocker, Singularity, comparison, metrics.

---

1 DevOps (Development and Operation) refers in this context to continuous integration/continuous delivery (CI/CD).

**IBERGRID Contributions / 27**

# The CERN analysis preservation portal

**Author:** Lara Lloret Iglesias[1]

[1] *CSIC*

**Corresponding Author:** lloret@ifca.unican.es

The CERN analysis preservation portal (CAP) comprises a set of tools and services aiming to assist researchers in describing and preserving all the components of a physics analysis such as data,

software and computing environment. Together with the associated documentation, all these assets are kept in one place so that the analysis can be fully or partially reused even several years after the publication of the original scientific results. An experiment-specific submission and retrieval interface has been developed for the CMS collaboration. It integrates with the CMS internal analysis registry (CADI) to capture all analyses with basic information, complemented with a detailed submission form for full information. The CMS data aggregation system (DAS) is interfaced to the deposit form to assist in filling in exact dataset names used in the analysis to ensure searchability. Efforts are ongoing to describe physics content for an intelligent retrieval, and to interface with container solutions for full reproducibility for selected test cases.

**IBERGRID Contributions / 28**

## Using Cloud Computing and Open Data to Improve Knowledge in the Insurance Sector

**Authors:** Alfredo Ferrer[1]; David Íñiguez[1]; Gonzalo Ruiz[1]

[1] *Instituto Universitario de Investigación Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza*

**Corresponding Authors:** aferrer@bifi.es, david.iniguez@bifi.es, gruiz@bifi.es

We present a technology transfer project where Cloud Computing and Open Data play a crucial role. Our aim is to accurately and efficiently model data from the Spanish car insurance sector. Due to the vast amount of data and the complexity of the models, the use of Cloud Computing is needed to ensure not only an efficient but also a feasible implementation of the model. The system was deployed on the OpenStack cloud platform of our Institute and it is portable to other cloud services such as Amazon Web Services. In addition to the usage of cloud technologies we also benefit from Big Data tools such as TensorFlow, ElasticSearch, Kibana or Spark.

The insurance sector is an important and growing sector of the Spanish's economy, representing a 5.5% of the GDP in 2017. Our data comes primarily from the quote calculator Avant2 of the software company Codeoscopic. This calculator, allows insurance agents to evaluate a specific risk (vehicle, driver,...) with many insurance companies and get quotes for different modalities. However, the companies' quoting criteria is a black-box. Finding information about this underlying process could shed light on understanding the differences between companies or regions and, ultimately, improve the Avant2 platform. Nonetheless, the companies' quotes were not completely explained by using only direct variables associated with the risk. To overcome this hurdle, we will also nourish our model with geographical data such as climate conditions, traffic accidents or socio-economic variables. This information was collected from several open source portals. Once we incorporated the open data component we find a significant improvement on the model's accuracy compared to only using internal data.

**IBERGRID Contributions / 29**

## Orchestrated satellite data management

**Authors:** Daniel Garcia Diaz[1]; Fernando Aguilar Gómez[1]

[1] *IFCA*

**Corresponding Authors:** garciad@ifca.unican.es, aguilarf@ifca.unican.es

With the latest missions launched by ESA or NASA, such as Sentinel or Landsat, equipped with the latest technologies in multispectral sensors, we face an unprecedented amount of satellite data never reached before. Exploring the potential of this data with state-of-the-art Artificial Intelligence techniques such as "Deep Learning" could potentially change the way we understand the Earth system and how to protect its resources.

The eXtreme-DataCloud project (XDC), under the umbrella of the H2020 programme, aims at developing a scalable environment for data management and computing, addressing the problems of the growing data volume and focused in providing a complete framework for research communities through the European Open Science Cloud. The target of this project is to integrate different services and tools based on Cloud Computing to manage Big Data sources, and Use Cases from diverse disciplines are represented. One of the goals of the project is to deal with extremely large datasets, including diverse data and metadata types, formats and standards that enable the automatic integration of Big Data.

In order to interoperate those big data sources, the XDC LifeWatch ERIC Use Case proposes a Virtual Research Environment (VRE) deployed on the Cloud that allow the users to preprocess the satellite data to obtain valuable information about the water quality of lakes and reservoirs without the need of using local resources as well as hiding the complexity behind. The architecture of this virtual environment consists of different Docker containers that run automatically with a common distributed storage system (Onedata) capable of storing the data with associated metadata that facilitate the discovery. The workflow of the VRE to preprocess the satellite data is manage by the INDIGO PaaS Orchestrator.

This presentation will describe the architectural design of the VRE and the different components (Jupyter interface, docker deployment for data preprocesing, modelling, etc.) as well as details on how this cloud-based approach can be adopted to many other cases.

**IBERGRID Contributions / 30**

# TRAFAIR: Understanding Traffic Flow to Improve Air Quality

**Authors:** Cecilia Grela Llerena[1]; Laura Po[2]; Federica Rollo[2]; Ríos Viqueira José Ramón[3]; Raquel Trillo Lado[4]; Alessandro Bigi[2]; Javier Cacheiro López[1]; Paolo Nesi[5]

[1] *Galicia Supercomputing Centre (CESGA)*

[2] *'Enzo Ferrari' Engineering Department*

[3] *Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS)*

[4] *Aragon Institute of Engineering Research (I3A)*

[5] *DISIT, University of Florence*

**Corresponding Authors:** cecilia.grela.llerena@cesga.gal, paolo.nesi@unifi.it, jlopez@cesga.es, alessandro.bigi@unimore.it, federica.rollo@unimore.it, jrr.viqueira@usc.es, raqueltl@unizar.es, laura.po@unimore.it

Road traffic is among the main sources of air pollution, and taking into account that air pollution causes 400 000 deaths per year, making it first environmental cause of premature death in Europe, environmental impacts of traffic are of major concern throughout many European metropolitan areas.

In February 2017, the European Commission warned five countries, among which Spain and Italy, of continued air pollution breaches. In this context, public administrations and citizens suffer from the lack of comprehensive and fast tools to estimate the level of pollution on an urban scale resulting from varying traffic flow conditions that would allow optimizing control strategies and increase air quality awareness.

TRAFAIR project surged from this premise, it brings together 9 partners from two European countries (Italy and Spain) to develop innovative and sustainable services combining air quality, weather conditions, and traffic flows data to produce new information for the benefit of citizens and government decision-makers. The project started in November 2018 and will lasts two years.

The TRAFAIR project aims at achieving four main results:
1) Definition of a standard set of metadata (based and extending the ones adopted at European level and defined by FAIRMODE) able to represent urban air quality maps.
2) Provision of real time estimation on air pollution in the city on an urban scale (using a set of

low-cost air quality sensors, and combined them with measurements by the regulatory air quality stations in order to build an informative map of the different levels of pollution in the urban areas).
3) Development of a service for prediction of urban air quality based on weather forecast and traffic flows. This service make use of open source and HPC technologies in order to compute the estimation of the diffusion of pollutants in the urban area. .
4) Publication of an open dataset describing the urban air quality maps and the prediction maps in 6 European cities of different size on which the service will run for the duration of the project: Zaragoza (600000 inhabitants), Florence (382000), Modena (185000), Livorno (160000), Santiago de Compostela (95000), Pisa (9.000). These datasets (including metadata) will be published on catalogs harvested by the European Data Portal.

The project is co-financed by the European Commission under the CEFTELECOM call on Open Data.

**IBERGRID Contributions / 31**

## Baseline criteria for achieving software quality within the European research ecosystem

**Authors:** Pablo Orviz Fernández[1]; Mário David[2]

[1] *IFCA-CSIC*

[2] *LIP*

**Corresponding Authors:** orviz@ifca.unican.es, david@lip.pt

Releasing the "A set of Common Software Quality Assurance Baseline Criteria for Research Projects" document (hereby referred to as "SQA baseline criteria") resulted from the need of filling up an uncovered gap in the European research software engineering ecosystem. This document sets a Software Quality Assurance (SQA) plan that maintains a pragmatic set of requirements, best practices and recommendations to drive an adequate development, timely delivery and reliable operation of the produced software assets within a research software development project.

The SQA baseline criteria covers the basic practices of making the source code open and accessible, pointing to the relevant open-source licenses and code hosting platforms. In what relates to source code management, it provides specific guidance in the usage of a change-based approach, by means of a version control system (VCS), that relies on a branching model to handle the addition of incoming new features or bug fixes, separating development and stable versions. Every relevant change in the code must be tested to avoid disruptions in the supported major branches or releases.

By following the aforementioned change-based approach, the SQA baseline criteria emphasizes the idea of acting at the early stages of the software lifecycle as the catalyst for maximizing the effectiveness of resolving issues (bugs, security flaws) with the lowest effort and cost. In this regard, the primary focus is put on the static analysis testing (such as unit/functional testing and vulnerability scanning), encouraging developers to have meaningful test cases that provide enough coverage of the system operation. At this stage, the readability and maintainability of the code are also essential quality requirements, achievable by making the source code compliant with a relevant programming language's style standard.

The documentation attached to the software is key to its adoption, and the SQA baseline criteria suggests that it be treated as code, through the use of markup languages and VCSs. Consequently, the documentation is versioned, with the capability of being rendered in multiple online documentation repositories. As the last requirement in the described change-based approach, a human-based review shall be performed in order to consider a set of aspects that cannot be assessed automatically, such as the change suitability or the understandability of the documentation.

The best practices at later stages include the interoperability assessment by the execution of integration tests that ensure the operation with external components, open standards and protocols. A

further security analysis is also performed at this stage, by checking common security flaws, thus covering two of the fundamental pillars of the dynamic analysis of the software.

The SQA baseline criteria as here presented has been elaborated based on the first-hand experiences of several European-funded software development projects. It is actively maintained (currently on version 2.0), online available, and open to collaboration and discussion. The aim is to keep improving and extending the document in order to consolidate it as a reference point for future research projects that involve development of software.

**IBERGRID Contributions / 32**

## Orchestration of optimized GPU containers using a cloud management platform in a hybrid environment

**Authors:** César Gómez-Martín[1]; Javier Alonso-Gómez[1]; Guillermo Díaz[2]; Alfonso Pardo[2]; José M. Franco[2]

[1] *Atrio*

[2] *CETA-Ciemat*

**Corresponding Authors:** javier.alonso@atrio.io, cesar@atrio.io, alfonso.pardo@ciemat.es, guillermo.diaz@ciemat.es, josemiguel.franco@ciemat.es

Nowadays, with the advent of container technologies like Docker or Singularity there is no need to have applications installed on any scientific resource; instead, they can be encapsulated inside a container to run in any operating system, CPU architecture, interconnect, and even leverage specific hardware accelerators like GPUs.

With a proper container definition, hardware or GNU/Linux distro-related incompatibilities can be seamlessly avoided, while retaining native or near-native performance depending on the host and the container configuration.

With ATRIO Composable Cloud (ACC) we provide a hardware and resource manager agnostic computing platform that is able to orchestrate any workload and use the bare-metal clusters in CETA-Ciemat (Slurm, OpenStack Nova or Ironic deployments), public cloud service providers (AWS, Azure, Google Cloud, etc.) and potentially any compute resource, like the European Open Science Cloud. In this poster, we show how ACC can orchestrate optimized containers using a variety of computing resources and deployments provided by CETA-Ciemat and some public cloud providers. We also demonstrate how simple is to deploy a machine learning workflow using Tensorflow with MPI, CUDA and low-latency interconnect (Infiniband) capabilities across multiple heterogeneous clusters.

**IBERGRID Contributions / 33**

## Integration of Apache Mesos over GPUs resources in the DEEP Hybrid DataCloud project

**Author:** Aida Palacio Hoz[1]

[1] *IFCA*

**Corresponding Author:** aidaph@ifca.unican.es

DEEP Hybrid DataCloud project was proposed with the necessity to support different amount of intensive computing techniques over specialized hardware, like HPC or GPUs. The project focus on the integration of this specialized, and expensive, hardware under a Cloud Platform as OpenStack that can be used on-demand by researchers of different areas. Within this project, a set of building blocks

whose solution is called “DEEP as a Service” was implemented to make the application deployment more easier for the user. For this development, it is necessary to provide the researchers with access to these technologies as friendly but powerful services able to exploit very large datasets.

On the one hand we have the gpu resources and on the other hand the users and their applications to run over those resources. DEEP needs to provide a service that controls how users can use those resources in an efficiently way. Although there are multiple technologies that address this problem as a queuing system, Apache Mesos has been developed to do it in an effective and controlled manner. Apache Mesos is a technology that abstracts the resources from different machines, like cpu, gpu, ram and storage to provide a scheduling and distributed system across the whole cloud environment. Mesos is easily to deploy over cpu-based systems but gpu needs a more tricky configuration as it is shown in this solution. As an added value, this solution provides an apache2 configuration for authenticate users from different communities by the current AAI service in DEEP.

The proposed presentation will show the design and deployment of Mesos to work over gpus resources inside the Deep Hybrid DataCloud Project scope.

**IBERGRID Contributions / 34**

## Past and Future Challenges for Distributed Computing at the ATLAS Experiment on the Iberian Peninsula

**Author:** Helmut Wolters[1]

**Co-authors:** Santiago González de la Hoz [2]; Carlos Acosta-Silva [3]; Javier Aparisi Pozo [2]; Mário David [1]; Jose Del Peso [4]; Álvaro Fernández Casani [2]; José Flix Molina [5]; Esteban Fullana Torregrosa [2]; Carlos García Montoro [2]; Jorge Gomes [1]; Julio Lozano Bahilo [2]; João Paulo Martins [1]; Gonzalo Merino [6]; Almudena del Rocio Montiel [4]; Andreu Pacheco Pages [3]; João Pina [1]; Javier Sánchez Martínez [2]; José Salt [2]; Aresh Vedaee [3]

[1] *LIP*

[2] *IFIC*

[3] *IFAE,PIC*

[4] *UAM*

[5] *PIC,CIEMAT*

[6] *CIEMAT,PIC*

**Corresponding Authors:** jorge@lip.pt, jpina@lip.pt, david@lip.pt, martinsj@lip.pt, helmut@coimbra.lip.pt

ATLAS is one of the big detector experiments at the Large Hadron Collider (LHC) at CERN. The LHC is in scheduled shutdown until end of 2020 for upgrading both collider and detectors which also provides new challenges on the ATLAS distributed computing (ADC). The higher luminosity in the next run will increase significantly data rate and storage needs, and also higher efficiency in the data treatment will be required. We have a longer time scale, the next scheduled upgrade to the High-Luminosity LHC that is foreseen to start during 2026 with an even bigger impact needs long-time preparation, both on worldwide storage and computing infra-structure, and on software tools.
The Iberian ATLAS Tier-1 and Tier-2s in Spain and Portugal form one regional component of the worldwide ADC infra-structure. They have more than 15 years of experience in the deployment and development of LHC computing components and their successful operations. The sites are already actively participating in, and even coordinating, emerging R&D computing activities developing the new computing models needed in the LHC Run3 and HL-LHC periods.
In this contribution, we present details on these development works such as

- HPC computing resources to execute ATLAS simulation workflows;
- the development of new techniques to improve efficiency in a cost-effective way, such as storage and CPU federations;

- recent developments of new monitoring tools that allow a more efficient control of the worldwide computing and storage operations;
- and improvements in Data Organization, Management and Access through storage consolidations ("data-lakes"), the use of data Caches, and improving experiment data catalogues, like Event Index.

The design and deployment of novel analysis facilities using GPUs together with CPUs and techniques like Machine Learning will also be presented.
We present the status of the Iberian ATLAS Tier-1 and Tier-2 sites, taking into account the national perspectives and how they can continue contributing to the significant R&D in computing by evaluating different models and for improving performance of computing and data storage capacity in the LHC High Luminosity era.

**IBERGRID Contributions / 35**

## A Data Science framework in the INCD

**Authors:** António Antunes[1]; Tiago Martins[1]; José Barateiro[1]; Anabela Oliveira[1]; Alberto Azevedo[1]

[1] *Laboratório Nacional de Engenharia Civil*

**Corresponding Authors:** aantunes@lnec.pt, tmmartins@lnec.pt, jbarateiro@lnec.pt, aazevedo@lnec.pt, aoliveira@lnec.pt

INCD - National Distributed Computing Infrastructure is a Portuguese digital infrastructure designed to support the national scientific community, providing computing and storage services to the national scientific and academic community in all areas of knowledge. LNEC – National Laboratory for Civil Engineering is one of the partners that collaborate in this initiative, developing use cases that take advantage from the available infrastructure. This work reports a Data Science framework based on Conda that was developed as part of this collaboration. The use of this framework allows researchers to benefit from the INCD infrastructure, running their research scripts, using the several Conda packages available, including Jupyter Notebook. To showcase the framework, two case studies were implemented, demonstrating the use of Machine Learning algorithms applied to data generate from the dam safety monitoring systems.
The first case study presents a prediction setting, implemented in Python, in which Multiple Linear Regression (MLR) and Neural Networks (NN) are trained and used to predict dam behavior in manually collected data. Environmental variables are used as predictors for both the MLR and the NN. Both predictions are evaluated and compared, also using the develop framework. Note that such predictions heavily depend on the specific properties of each data set. Thus, the capabilities of this environment on top of INCD infrastructure enable a flexible adaptation of each prediction that can be easily tuned to each specific case.
A classification task is proposed in the second case study, implemented in Python and R, using the DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm to identify outliers in automatically collected data from sensors installed on Portuguese dams. Together with the dam response variable, environmental variables are used to obtain the clusters and detect outliers. Afterwards, PCA (Principal Component Analysis) is used to obtain a 2D plot to visualize outliers identified by the DBSCAN.
Other pieces of research were also developed using this framework, including the use of Deep Learning (more specifically, Recurrent NN) to improve prediction of dam behavior, using Keras and TensorFlow, which benefited from the INCD infrastructure for improved computation times.
Finally, it is important to remark that framework was recently presented to several LNEC researchers and was received with a large interest, with most of the researchers already starting to use INCD to create and run their scripts in a cloud environment.

**IBERGRID Contributions / 36**

## Machine Learning Pipelines on Medical Imaging

**Authors:** Walter Dos Santos[1]; Wagner Meira Jr.[1]; Eduardo Camacho-Ramos[2]; Ana Jiménez-Pastor[2]; Ángel Alberich-Bayarri[2]; Ignacio Blanquer Espert[3]

[1] *Universidade Federal de Minas Gerais*

[2] *QUIBIM*

[3] *Universitat Politècnica de València*

**Corresponding Authors:** meira@dcc.ufmg.br, iblanque@dsic.upv.es, angel@quibim.com, anajimenez@quibim.com, educamacho@quibim.com, walter@dcc.ufmg.br

The use of Artificial Intelligence (AI) over medical data allows the extraction of features associated to the disease from medical images using data-characterisation and modelling algorithms. The use of advanced machine learning algorithms is changing the way image processing is performed, evolving from analytic solutions to models built up with supervised training techniques working in complex Convolutional Neural Network (CNN) architectures. However, advanced AI techniques require a deep understanding of the behaviour of the model and non-trivial programming skills. This limits the application of AI to researchers who have a deep understanding of the medical problem but lack from those specific technical skills.
In this work, we will compare an application for the automatic diagnosis of Rheumatic Heart Disease (RHD) from echocardio videos on children, implemented using Keras with an equivalent application deployed using a machine learning workflow system (LEMONADE).
The processing pipeline requires 7 steps: frame splitting, which splits a video into frames; automatic classification into doppler and anatomical images by color inspection (only doppler images are used during the rest of the pipeline); color-based segmentation through k-means clustering; image preprocessing and view classification by using a CNN; first- and second-order texture analysis and blood-flow velocity calculation; z-score features normalization; and classification of extracted features through machine learning techniques into RHD positive or healthy studies.
The implementation in Keras uses the pre-trained models for the classification of the views within the estimation of the RHD. All the components are delivered as containers, facilitating their distribution and the integration of new components in LEMONADE.
The processing backend is a Kubernetes cluster provided of GPU nodes attached through PCI passthrough to the Virtual Machines and the containers. This way there is no penalty on the usage of the GPUs from the applications. Data are stored directly on a persistent storage object exported through an SSH server. As communications are encrypted, data access is measured separately.

**IBERGRID Contributions / 37**

## Data Science in High Energy Physics

**Authors:** Rute Pedro[1]; Tiago Vale[2]; Nuno Castro[3]; Guilherme Milhano[2]; Crispim Romao[None]

[1] *LIP/FCUL*

[2] *LIP*

[3] *LIP, and University of Minho*

**Corresponding Authors:** rute@lip.pt, nuno.castro@cern.ch, tvale@lip.pt, gmilhano@lip.pt

High Energy Physics is a big data task that requires modern data science tools for storage, processing and analyzes. In this contribution we aim to overview the applications of machine learning, namely the modern deep learning approach, to aid research in collider physics and related topics. More specifically we will show how Convolutional Neural Networks can help us learn about new observables for jet physics and other Artificial Neural Networks are becoming the new the paradigm for data analysis at the Large Hadron Collider. Due to the complexity of the task and volume of the data used these neural networks are implemented in Keras using Tensorflow and trained on high performant Graphical Processing Units.

**IBERGRID Contributions / 38**

## DEEP-Hybrid Datacloud: a project summary

**Author:** Alvaro Lopez Garcia[1]

[1] *IFCA-CSIC*

**Corresponding Author:** aloga@ifca.unican.es

The DEEP-Hybrid-DataCloud project researches on intensive computing techniques such as deep learning, that require specialized GPU hardware to explore very large datasets, through a hybrid-cloud approach that enables the access to such resources. DEEP is built on User-centric policy, i.e. we understand the needs of our user communities and help them to combine their services in a way that encapsulates technical details the end user does not have to deal with. DEEP takes care to support users of different levels of experience by providing different integration paths. We show our current solutions to the problem, which among others include the Open Catalog for deep learning applications, DEEP-as-a-Service API for providing web access to machine learning models, CI/CD pipeline for user applications, Testbed resources. We also present our use-cases tackling various problems by means of deep learning and serving to demonstrate usefulness and scalability of our approach.

**IBERGRID Contributions / 39**

## Distributed Computing at the CMS Experiment

**Author:** Diogo de bastos[None]

**Corresponding Author:** dbastos@lip.pt

Being one of the largest international scientific collaborations, CMS faces many challenges. To serve the computational needs of every researcher working around the world within the Collaboration, CMS relies on distributed computing technology for both computing power and data storage. The Large Hadron Collider (LHC) schedule alternates between data-taking periods and long shutdowns for maintenance and upgrades. Currently on the Long Shutdown 2, the CMS detector is being upgraded. Run 3 is scheduled to start in 2021 with an increase in luminosity. These two facts combined will pose new challenges for LHC's distributed computing and data storage infrastructure called the Worldwide LHC Computing Grid (WLCG). Aiming to increase the nominal luminosity by a factor of 5-7 a major upgrade to the LHC is expected to start after run 3 in 2026 called the High Luminosity Large Hadron Collider (HL-LHC). Preparations for this upgrade have already started.

As a member of the WLCG collaboration, Portugal has pledged to contribute to CMS Tier-2 sites with CPU and storage responsibilities.

In this talk, we will present a brief overview of the involvement of the Portuguese group in the CMS experiment. Starting from the physics analysis being done at LIP CMS group to the computational needs we foresee for the next 10 years. We will cover the tools used for our physics analyses, our computational needs, the Portuguese role in the Tier-2 management and how we are going to address the expected necessities in the next 10 years.

**IBERGRID Contributions / 40**

## Understanding and forecasting the Portuguese marine environment: the activity of Instituto Hidrográfico in the area of physical oceanography.

**Author:** Joao Vitorino[1]

[1] *IH*

**Corresponding Author:** joao.vitorino@hidrografico.pt

Instituto Hidrografico is a Portuguese State Laboratory founded in 1960 which have as main mission the monitoring and study of the marine environment in order to support the Portuguese Navy and to contribute to the national development in the areas of Marine Sciences and Marine Technologies. The activity of Instituto Hidrografico covers domains such as hydrography/cartography, physical oceanography, marine geology, marine chemistry and pollution and safety to navigation. In this contribution we focus on the area of physical oceanography view as an excellent example of the commitments, challenges and opportunities faced today by Instituto Hidrografico. Central in the activity developed in this area is the operation of a large real-time monitoring infrastructure covering the Portuguese marine area, which includes observing systems installed both in land as well as offshore the coast. These different systems generate a large flow of data that is received daily at Instituto Hidrografico and from here disseminated to different users. In addition to this permanent effort of observation other more time-limited programs of observations are conducted, namely during multidisciplinary surveys onboard hydrographic vessels. The observation activity is complemented and extended by numerical modelling activities. Numerical models are used at Instituto Hidrografico to provide in-depth understanding of the oceanographic processes, to allow that a comprehensive 3(4)D picture of the marine environment be built from the observations and to forecast the future evolution of oceanographic conditions from the knowledge of the present state of the ocean. These different areas of activity are supported in a number of infrastructures installed at Instituto Hidrografico, namely computer clusters for parallel computing. They all have benefit from the inclusion of Instituto Hidrografico as partner in different national and Europeans projects such as (among the most recent) the JERICO-NEXT (H2020-INFRAIA), MARISK (INTERREG) or MYCOAST (EU INTERREG Atlantic Area).

**IBERGRID Contributions / 41**

## Object storage for climate data storage and analytics

**Authors:** Ezequiel Cimadevilla Alvarez[1]; Aida Palacio[2]; Antonio Cofiño[3]

[1] *Santander Meteorology Group (unican)*

[2] *IFCA*

[3] *Santander Meteo Group (UNICAN)*

**Corresponding Authors:** ezequiel.cimadevilla@unican.es, aidaph@ifca.unican.es, antonio.cofino@unican.es

Data analysis in climate has been traditionally done in two different environments, local workstations and HPC infrastructures. Local workstations provide a non scalable environment in which data analysis is restricted to small datasets that are previously downloaded. On the other hand, HPC infrastructures provide high computation capabilities by making use of parallel file systems and libraries that allow to scale data analysis.

Parallel file systems found in HPC show scalability limitations due to constraints of the POSIX standard, which favors consistency of files while penalizing scalability. Object storage consists of a new storage system that tries to favor scalability instead of consistency and is usually provided by commercial cloud storage providers, such as Amazon S3 or Google Cloud Storage.

NetCDF, the standard library for climate data, actually requires files to be stored in a file system, although work is currently being carried out to allow storage on object stores. In this work we show how these new object storage systems can be combined with Python libraries, such as xarray and Dask for distributed analysis and Zarr for data storage in object stores, that allow computations to be easily parallelized without scalability restrictions.

**IBERGRID Contributions / 42**

## SOCIB regional ocean observing and forecasting infrastructure in the Western Mediterranean Sea

**Author:** Baptiste Mourre[1]

[1] *SOCIB - CSIC*

**Corresponding Author:** bmourre@socib.es

SOCIB (Balearic Islands Coastal Observing and Forecasting System, www.socib.es) is a coastal ocean observing and forecasting infrastructure located in the Western Mediterranean Sea. SOCIB collects and distributes data from near-shore to open ocean through the operation of multi-platform observing systems from fixed moorings, drifting buoys, research vessel, gliders, HF radar, animal tracking systems and beach monitoring stations. It provides free and quality-controlled observations and products to address both science and society needs.

SOCIB operates three ocean prediction systems aiming to predict the short-term evolution of (1) ocean temperature, salinity, sea level and currents (2) waves and (3) meteotsunamis. Their outputs are being disseminated on the web and integrated in specific SOCIB products and services tailored to the needs of specific sectors and end-users.

This presentation will provide an overview of the main components of SOCIB, from observations to prediction systems and applications, also including collaborative projects with Iberian partners.

**Plenary Session on Environmental Sciences / 43**

## Understanding and forecasting the Portuguese marine environment: the activity of Instituto Hidrográfico in the area of physical oceanography.

Instituto Hidrografico is a Portuguese State Laboratory founded in 1960 which have as main mission the monitoring and study of the marine environment in order to support the Portuguese Navy and to contribute to the national development in the areas of Marine Sciences and Marine Technologies. The activity of Instituto Hidrografico covers domains such as hydrography/cartography, physical oceanography, marine geology, marine chemistry and pollution and safety to navigation. In this contribution we focus on the area of physical oceanography view as an excellent example of the commitments, challenges and opportunities faced today by Instituto Hidrografico. Central in the activity developed in this area is the operation of a large real-time monitoring infrastructure covering the Portuguese marine area, which includes observing systems installed both in land as well as offshore the coast. These different systems generate a large flow of data that is received daily at Instituto Hidrografico and from here disseminated to different users. In addition to this permanent effort of observation other more time-limited programs of observations are conducted, namely during multidisciplinary surveys onboard hydrographic vessels. The observation activity is complemented and extended by numerical modelling activities. Numerical models are used at Instituto Hidrografico to provide in-depth understanding of the oceanographic processes, to allow that a comprehensive 3(4)D picture of the marine environment be built from the observations and to forecast the future evolution of oceanographic conditions from the knowledge of the present state of the ocean. These different areas of activity are supported in a number of infrastructures installed at Instituto Hidrografico, namely computer clusters for parallel computing. They all have benefit from the inclusion of Instituto Hidrografico as partner in different national and Europeans projects such as (among the most recent) the JERICO-NEXT (H2020-INFRAIA), MARISK (INTERREG) or MYCOAST (EU INTERREG Atlantic Area).

**Plenary Session on Environmental Sciences / 44**

## SOCIB regional ocean observing and forecasting infrastructure in the Western Mediterranean Sea

SOCIB (Balearic Islands Coastal Observing and Forecasting System, www.socib.es) is a coastal ocean observing and forecasting infrastructure located in the Western Mediterranean Sea. SOCIB collects and distributes data from near-shore to open ocean through the operation of multi-platform observing systems from fixed moorings, drifting buoys, research vessel, gliders, HF radar, animal tracking systems and beach monitoring stations. It provides free and quality-controlled observations and products to address both science and society needs.

SOCIB operates three ocean prediction systems aiming to predict the short-term evolution of (1) ocean temperature, salinity, sea level and currents (2) waves and (3) meteotsunamis. Their outputs are being disseminated on the web and integrated in specific SOCIB products and services tailored to the needs of specific sectors and end-users.

This presentation will provide an overview of the main components of SOCIB, from observations to prediction systems and applications, also including collaborative projects with Iberian partners.

**Plenary Session on Environmental Sciences / 45**

## Computational engineering services for all: LNEC experience as a INCD/IBERGRID/EOSC user

The computational resources necessary to address major environmental scientific questions are seldom available in-house, making shared e-infrastructures a well-suited medium for performing complex model simulations, analyzing large datasets and applying decision support tools. Despite this potential, the technical expertise required to use these computational resources and to build products on top of them is very specialized and requires a combination of environmental scientists and computer science engineers for their development and maintenance.
In the scope of the Portuguese Infrastructures Roadmap and of two H2020 European Open Science Cloud e-infrastructures projects, several e-services dedicated to environmental sciences have been developed by LNEC and its partners and made freely available to promote the work of environmental scientists and engineers. These services encapsulate several state-of-the-art numerical models and data analysis tools, and are offered through dedicated, user-friendly Web apps. These tools hide the complexity of e-infrastructures resources allocation from the user and simplify the application of the modeling and data components.
This presentation presents two of these services in detail:
• OPENCoastS, a service that assembles on-demand circulation forecast systems for user-selected coastal areas and keeps them running operationally for a period defined by the user, using INCD and IFCA computational resources.
• WorSiCa (Water mOnitoRing SentInel Cloud platform), a service that integrates remote sensing and in-situ data for the determination of water presence in coastal and inland areas, applicable to a range of purposes from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks.
The OPENCoastS service is based on the application of the modeling suite SCHISM and generates daily forecasts of water levels and vertically averaged velocities over the region of interest for 48 hours, based on numerical simulations of the relevant physical processes.
WorSiCa is a one-stop-shop service to provide access to customized remote sensing services based on Copernicus data, currently applied to the detection of the coastal water land interface and the inland water detection (for large water infrastructure leak detection).

**Plenary Session on Environmental Sciences** / **46**

## Improving access and use of GBIF through infrastructure cooperation at the Iberian level

The Global Biodiversity Information Facility (GBIF) is a global government-level effort to mobilise and make freely available online primary biodiversity data for all biological groups. Through GBIF, more than 1.3 billion records are currently available globally and 39 million for the Iberian Peninsula. Both Portugal and Spain implemented national data portals to facilitate users' access to biodiversity data in full context and advanced ways, not available at the global level.

The main drivers of biodiversity distribution are related to environmental and climatic factors. Species occurrences are not constrained by political borders. Therefore, access to biodiversity data for scientific and management purposes should be possible under a biogeographic context, enabling analysis of information in ecologically meaningful scopes. Moreover, information systems and technological platforms should promote cross-border cooperation, so that species distribution modelling, species invasion, red listing and other conservation efforts can be seamless performed by researchers and users at the iberian level.

Many GBIF participants have adopted the opensource Atlas of Living Australia (ALA) platform, creating the community Living Atlases (LA), in which both Portugal and Spain participate. In these countries, the national portals are supported by cloud computing services provided by IBERGRID partners, INCD and IFCA, respectively. These portals have been operating for more than three years, providing thousands of accesses annually. The LA architecture is modular, including several APIs built on top of an infrastructure layer of databases (Cassandra, MySQL), file storage and indexes (SOLR). It is on the creation and configuration of this infrastructure layer that the cloud computing excels, particularly in testing and updating environments.

The web applications of LA platform provide information integration, allowing visualization of data on lists, maps, images formats and metadata. It is possible to create online reports of species lists based on localities or areas. Using spatial modules, biodiversity information can be crossed with geographic or spatial and environmental data, providing even more detailed reports. There are also analysis tools to perform species distribution modeling, red list assessments and other biodiversity-based analysis.

A single infrastructure of LA can support different portals using the hub module. In this way, it is possible to enable a thematic, an institutional or a regional portal. In this presentation, we will explore how the platform can be extended to share biodiversity data across Portugal and Spain, providing biogeographic-based facets that allows searches and analysis, without breaks due to administrative borders. We will also discuss how cloud-based services based on the grid computing community can facilitate this integration at the Iberian level, enabling also redundancy of security and availability of service. This shared vision between GBIF Portugal and GBIF Spain, in the scope of the national infrastructures PORBIOTA and LifeWatch-ES, may contribute to a better support to research studies and natural resource management at the Iberian level.

**Research Communities & EOSC** / **47**

## Computational challenges related to IFMIF and DONES facilities.

Following ITER, DEMO reactor is expected to demonstrate the feasibility of safe, environmentally friendly and economically viable fusion power generation. During operation of DEMO, the materials will be exposed to a particular hostile environment as a consequence of the energetic neutrons created by fusion reactions in the plasma. The level of damage expected in fusion conditions is such that the performance of materials and components under these extreme irradiation conditions is unknown. One of the central objectives of the fusion materials program is to identify innovative materials development routes, using scientific

understanding and knowledge of how materials properties evolve and change in the operating environment of a fusion power plant.
In this respect, IFMIF is considered as one of the main pillars in the international fusion program. Its double deuteron beam 125 mA each will produce enough rate of damage behind the lithium target to make available in a few years information on materials damage at DEMO relevant doses. On the other hand, DONES (DEMO Oriented Neutron Source) has been conceived as a simplified IFMIF-like plant to provide in a reduced time scale and with a reduced budget – both compared to IFMIF- the basic information on materials damage.
Although both facilities are designed to provide experimental data on how the material properties change under energetic neutron irradiation, the design of experiments to be carried out to test materials implies various computational challenges. During our talk we shall review the different computational fields associated to IFMIF and DONES facilities, such as beam dynamics, neutronic transport, calculation of collision cascades and the simulation of the microstructure evolution in the irradiated materials.

**e-Infrastructure Plenaries** / **48**

## RDA Spain

**Scientific Data Repositories: a National perspective** / **49**

## European Data Incubator: fostering Big Data and AI driven economy in Europe

**Scientific Data Repositories: a National perspective** / **50**

## (FAIR4HEALTH presentation)

**Scientific Data Repositories: a National perspective** / **51**

## Big data, big responsibility: data lineage management with template for reproducible scientific papers

**Scientific Data Repositories: a National perspective** / **52**

## Research Data repositories at CSIC

**Scientific Data Repositories: a National perspective** / **53**

## The role of the German Council for Information Infrastructures

**Scientific Data Repositories: a National perspective / 54**

## Biodiversity Data repositories in Poland

**Plenary Session on Environmental Sciences / 55**

## LifeWatch ERIC: Consolidating synergies among Iberian e-Biodiversity communities through IBERGRID-IBERLIFE & EOSC (Synergy) initiatives

**Scientific Data Repositories: a National perspective / 56**

## DANS approach to Data repositories in the Netherlands

**Scientific Data Repositories: a National perspective / 57**

## FCT roadmap for scientific data repositories

**e-Infrastructure Plenaries / 58**

## Closing

**Corresponding Author:** christian.cuciniello@ec.europa.eu

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 59**

## Using OpenStack Cloud infrastructures

**Corresponding Author:** david@lip.pt

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 60**

## Performing computations using Docker containers in interactive and batch systems, in Grids, Cloud and HPC systems

**Corresponding Author:** jorge@lip.pt

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 61**

## Basic tutorial on Event-Driven computing applied to data processing

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 62**

## Development of basic Serverless systems using Docker containers and Jupyter Notebooks

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 63**

## DEEP-Hybrid Data Cloud: Deep learning tools

**Tutorial: Tools for exploiting Advanced Digital InfrastructuresInnovative Computing and Data Processing tools for Researchers / 64**

## Usage of Open Source cloud Platforms as a Service to compose, deploy and manage cloud services (Alien4Cloud and INDIGO-Datacloud platform)

**Evolution of production software tools in EOSC / 65**

## Welcome

**Corresponding Authors:** cristina.aiftimiei@cnaf.infn.it, david@lip.pt

Presenting the goal of session and contributors

**Evolution of production software tools in EOSC / 66**

## Cloud IaaS/PaaS and EGI Notebook service

This presentation will provide an overview of the ongoing and new developments coming to 3 of the EGI computing services: Cloud Compute — which offers a federated multi-cloud IaaS — , Cloud Container Compute — which offers a Kubernetes-based platform for running docker applications — and Notebooks, a completely managed interactive computing service based on Jupyter.

**Evolution of production software tools in EOSC / 67**

## Federated data management requirements and technical roadmap

The presentation provides an overview of the requirements gathered during the Data Management Workshop where XDC, ESCAPE and EGI met three important user communities to design with them some Research Infrastructure specific solutions and pilot activities. After this the EGI data-related services and their status are presented and it eventually looks forward presenting some scouting activities highlighting solutions that could complement and augment the EGI service offering.

**Evolution of production software tools in EOSC / 68**

## Check-in technical roadmap and RCauth status and plans

The EGI Check-in service is an Identity and Access Management solution that makes it easy to secure access to services and resources. Check-in is one of the enabling services for the EOSC-hub AAI following the architectural and policy recommendations defined in the AARC project. Through Check-in, users are able to authenticate with the credentials provided by the IdP of their Home Organisation (e.g. via eduGAIN), as well as using social identity providers, or other selected external identity providers. Check-in provides an intuitive interface for communities to manage their users and their respective groups, roles and access rights. For communities operating their own group management system, Check-in has a comprehensive list of connectors that allows to integrate their systems as externally managed Attribute Authorities. The adoption of standards and open technologies, including SAML 2.0, OpenID Connect, and OAuth 2.0, facilitates integration with web-based services. Options to support non-web services, which traditionally relied on X509 certificates, are based around the concept of online authorities with attached credential stores, such as RCauth.eu with a tightly-coupled MyProxy server. Such techniques allow science gateways to obtain credentials on behalf of the end-user that can be used to directly authenticate to services. Another user-centric approach considers certificate proxies as opaque tokens that can be obtained from a credential store from the command-line using SSH authentication. The deployed RCauth.eu and MasterPortal service from AARC features both these capabilities and has been shown to work for the production EGI and WLCG environments. The currently-operational RCauth.eu is being re-engineered to allow for state consistency between a geographically distributed set of hosting sites. The presentation will provide an overview of the EGI Check-in technical roadmap and the evolution of the RCauth service towards a distributed deployment architecture.

**Evolution of production software tools in EOSC / 69**

## DEEP-Hybrid Datacloud: present and future

**Corresponding Author:** aloga@ifca.unican.es

The DEEP-Hybrid-DataCloud is providing a set of comprehensive services for machine learning and deep learning, allowing scientists to train, test, evaluate, share and exploit their models over distributed e-Infrastructures. New advancements, will be presented and described, future exploitation of the solutions proposed

**Evolution of production software tools in EOSC / 70**

## eXtreme DataCloud: present and future

**Corresponding Author:** daniele.cesini@cnaf.infn.it

The eXtreme DataCloud (XDC) project is aimed at developing data management services capable to cope with very large data resources allowing the future e-infrastructures to address the needs of the next generation extreme scale scientific experiments. Started in November 2017, XDC is combining the expertise of 8 large European research organisations, the project aims at developing scalable technologies for federating storage resources and managing data in highly distributed computing environments.
The state of the art of the developed solutions, together with the new advancements, will be presented and described during the session.

**Evolution of production software tools in EOSC / 71**

## EOSC-hub TCOM SQA area: status and future

**Corresponding Author:** jpina@lip.pt

Overview of the EOSC-hub Technology Committee (TCOM) work done for the Software Quality Assurance area with special focus on the EOSC-hub technical workshop that served as input for EOSC architecture and service roadmap.