IBERGRID 2018

Report of Contributions

https://indico.lip.pt/e/437

Contribution ID: **1** Type: **Presentation**

HPC and Cloud Resources for Running Mathematical Simulations Efficiently

Complex simulations that require large amounts of computational resources have typically run in dedicated supercomputers. However some parts of these simulations don´t perform well in these computers or don´t need these highly costs resources and can be executed on cheaper hardware. Moving some parts of these simulations out of the supercomputers and running them in smaller clusters or cloud resources can improve the time to results and reduce the costs of the simulations, providing also higher flexibility and ease of usage. Both HPC and Cloud resources benefit and empower users who need to perform complex simulations that normally only take advantage of the capabilities of one of these infrastructures. We propose the combined usage of these platforms by using an orchestrator to coordinate the exploitation of these systems and container technology to enable interoperability between them. Such solution provides simulations as a service in a transparent way for end-users and software developers, as well as improves the efficiency in HPC resources usage. It has been proven to work with different HPC and Cloud providers, including EOSC Hub.

Primary author: CARNERO, Javier (Atos Research & Innovation)

Co-authors: Mr SANDE, Victor (Cesga); Mr DIAZ, Pablo (Cesga); Mr NIETO, Francisco Javier (Atos); Mr FERNANDEZ, Carlos (Cesga)

Track Classification: R&D for computing services, networking, and data-driven science at the Iberian level.

Contribution ID: **2** Type: **Presentation**

Scipion on-demand service in the cloud

Scipion is an image processing framework used to obtain 3D maps of macromolecular complexes on Cryo Electron Microscopy. It has emerged as the solution offered by the Instruct Image Processing Center (I2PC), hosted by CNB-CSIC, to European scientists accessing the European Research Infrastructure for Structural Biology (Instruct).

Cryo-EM processing is very demanding in terms of computing resources requiring powerful servers and since recently the use of GPUs. Common desktop machines are clearly insufficient in computing capability and storage which could a problem for many scientists that might not have access to big servers or GPUs.

Cloud IaaS (Infrastructure as a Service) is a new form of accessing computing and storage resources on demand. To effectively use cloud infrastructures ScipionCloud was developed, resulting in a full installation of Scipion both in public and private clouds, accessible as public "images", that include all needed cryoEM software and just requires a Web browser to work as if it was a local desktop. These images are available in the EGI Applications Database and in AWS public AMIs catalogue.

We present here a new service for Instruct users that would allow them to process the data acquired at any of the high end Instruct Facilities -focusing this initial work in own cryo EM Facility at the I2PC- on a virtual machine in one of the IberGRID sites. In this first scenario we are now presenting, the machine itself is setup by I2PC staff, but as we advance in our development we envision the opening of a web portal accessible to I2PC users to do that.

Primary authors: Prof. CARAZO, Jose Maria (Centro Nacional de Biotecnologia - CSIC); Mrs DEL CANO, Laura (Centro Nacional de Biotecnologia - CSIC)

Track Classification: Development of applications for supporting User Communities in the context of the EOSC.

Contribution ID: **3** Type: **Presentation**

On-premises Serverless Container-aware ARchitectures

Changes in the programing paradigms over the last years have pushed Cloud providers, such as Amazon Web Services (AWS), to offer new service solutions to adapt to the user requirements. Cloud serverless computing has arisen to palliate the demand of executing small pieces of code in the Cloud without having to previously provision infrastructure resources. Several benefits from serverless services (such as AWS Lambda) are high scalability, ease of deployment and a fine-grained pay-per-use policy. However, those platforms also impose significant restrictions for the applications and the environment in which they are executed (e.g. runtime, permissions and environment constraints).

To address such issues we developed a framework and a methodology to create Serverless Container-aware ARchitectures (SCAR). SCAR, in combination with udocker (a tool to execute Docker containers in user space), allows to run Docker containers in AWS Lambda, thus enabling the user to execute customized runtime environments and bypass some of the limitations imposed by the provider. The SCAR framework also aims to ease the application deployment process and allows the user to create, in addition to the serverless functions, API endpoints and use function composition to model their applications.

Although SCAR presents increased benefits to the usage of serverless functions on the Cloud, there are some strict limitations imposed by Cloud providers that cannot be bypassed only by software (e.g. storage size limit, execution time limit). This fact, in combination with data restrictions usually present in scientific applications, like private data that cannot be uploaded to the Cloud, encouraged us to create the On-premises Serverless Container-aware ARchitectures (OSCAR) framework. OSCAR will preserve the benefits of SCAR but in an on-premises infrastructure. In combination with curated software such as the IM and EC3, OSCAR offers the users the possibility of deploying an elastic Kubernetes cluster with serverless functions support, together with automated data management from S3-like storage backends.

Primary authors: PÉREZ GONZÁLEZ, Alfonso (UPV - GRyCAP); Dr MOLTÓ, Germán (UPV); Dr CABALLER, Miguel (UPV); Dr CALATRAVA, Amanda (UPV)

Track Classification: Development of Innovative Software Services oriented to EOSC

Contribution ID: **4** Type: **Lightning talks**

A New Role for Supercomputing Centers in Open Science

In the last decades an exponential increase of the scientific and technical development in all the areas of science has become manifest, being more and more relevant the conflicts between the pure scientific advance of the society and the property of the researched knowledge. Particularly the brake that certain aspects of the established system suppose in the acquisition of some findings. The ever demanding scientific community aims for new platforms and services up to date in order to diminish the inconveniences that have arisen due this same growth. The guidelines provided by the Open Science (OS), despite being a concept already discuss for some time, set out the perfect framework for the creation of new applications and platforms in which the research centers might take over from the responsibility exercised so far by the publishers to allocate resources for disclosure and distribution without replace them. In particular Open Access (OA), Open Data (OD) and Open Methodologies (OM) are the guidelines that can fit the most in the current tasks performed by supercomputing and research centers where availability, reliability and security are concepts well implemented already and can be very powerful skills to them in order to step in the spotlight in this new framework which is OS. In the other hand, while universities and publishers can be tempted by recognition and self promotion at the time of deciding whether a job can benefit them or not, the public supercomputing centers, that are already providing services for both private and public projects, working in the same direction and as a set, could be considered more objectives. In this paper it is proposed a path to follow by the supercomputing centers, within the framework of European Open Science Cloud (EOSC), to share resources with the aim of providing an adequate infrastructure for the development of scientific research, adding to their current competences the ability to become neutral ground for scientific disclosure.

Primary author: Dr JIMÉNEZ, Luis Ignacio (Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS))

Co-authors: Mr CALLE-CANCHO, Jesús (Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS)); Dr CORTÉS-POLO, David (Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS)); Dr GONZÁLEZ-SÁNCHEZ, José Luis (Research, Technological Innovation and Supercomputing Center of Extremadura (CénitS))

Track Classification: Development of Innovative Software Services oriented to EOSC

Contribution ID: **5** Type: **Presentation**

e-Science services evolution at RedIRIS

This presentation will outline the services currently in production, and also under development, oriented to support the access of researchers in Spain to the Digital Infrastructures of the EOSC era.The purpose of the presentation is also collecting feedback from the participants.

Primary author: FUENTES, Antonio (RedIRIS)

Track Classification: R&D for computing services, networking, and data-driven science at the Iberian level.

Contribution ID: **6** Type: **Presentation**

Easy use of Distributed TensorFlow Training on supercomputing facilities.

Deep Learning is a powerful tool for science, industry and other sectors that benefits from large datasets and computing capacity during models'design and training phases. TensorFlow (TF) Google's Machine Learning API is one of the tools most widely used for developing and training such deep learning models. There is a wide range of possibilities to configure a deep learning model however find the optimal model architecture can be a highly demanding computing task. Moreover, when involved datasets are very large the computing requirements increase and training processes can take a lot of time and hinder the design cycle. One of the most powerful capabilities of TF is its distributed computing capabilities, allowing portions of the automatic generated graph to be calculated on different computing nodes, and speeding up the training process. Deployment of distributed TF is not a straightforward task and it presents several issues, mainly related with its use under the control of local resources management systems and the usage of the right resources. In order to allow CESGA users to adapt their own TF codes to take advantage of TF and Finis Terrae II distributed computing capabilities, a complete python Toolkit has been developed. This Toolkit deals with several tasks that are not relevant in the models design, but necessary for exploiting the distributed capabilities, hiding the underlying complexity to final users. Additionally, an example of a successful industrial case, based on the Fortissimo 2 project experiment "Cyber-Physical Laser Metal Deposition (CyPLAM)", that uses this Toolkit, is presented. Thanks to the TF distributed capabilities, the computing capability of Finis Terrae and the use of the developed Toolkit, the time needed for training the largest model of this industrial case has been decreased from 8 hours (non- distributed TF) to less than 20 minutes.

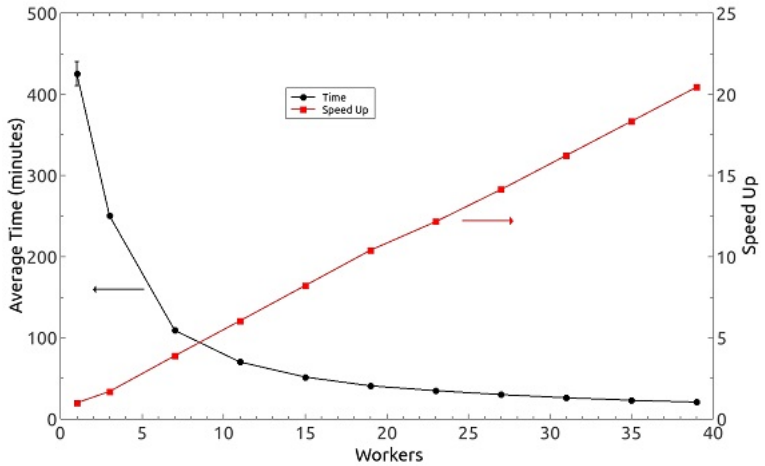

Figure 1: Training time (left axis) and Speed Up (right axis) vs number of tasks for Distributed TensorFlow training for a CyPLAM model.

Primary authors: Dr FERRO COSTAS, Gonzalo (CESGA); Mrs COTELO QUEIJO, Carmen (CESGA); Dr GÓMEZ TATO, Andrés (CESGA)

Track Classification: R&D for computing services, networking, and data-driven science at the Iberian level.

Contribution ID: **7** Type: **Presentation**

To Trust or Not To Trust, that is the question

Trust is a central issue confronting men and women in contemporary society. In fact, the most difficult thing to achieve in this world is trust. It can take years to win and only a matter of seconds to lose it. This is also applicable in a computing environment, where users need to trust computing services to process and manage their data. This implies a broad spectrum of properties to be accomplished, such as Security, Privacy, Coherence, Isolation, Stability, Fairness, Transparency and Dependability.
Adaptive, Trustworthy, Manageable, Orchestrated, Secure Privacy-assuring Hybrid, Ecosystem for REsilient Cloud Computing1 (hereinafter "ATMOSPHERE") is a European-Brazilian collaboration project aiming at measuring and improving the different trustworthiness dimensions of data analytics applications running on the cloud. To achieve trustworthy cloud computing services on a federated environment, ATMOSPHERE focuses on providing four components: i) a dynamically reconfigurable hybrid federated VM and container platform, to provide isolation, high-availability, Quality of Service (QoS) and flexibility; ii) Trustworthy Distributed Data Management services that maximise privacy when accessing and processing sensitive data; iii) Trustworthy Distributed Data Processing services to build up and deploy adaptive applications for Data Analytics, providing high-level trustworthiness metrics for computing fairness and explainability properties; and iv) a Trustworthy monitoring and assessment platform, to compute trustworthiness measures from the metrics provided by the different layers.
In this lightning session, we will focus our discussion in the integration of the federated cloud platform with the Trustworthy monitoring and assessment platform, in order to provide isolation, stability and Quality of Service performance guarantees. The cloud platform will enable the dynamic reconfiguration of resource allocation to applications running on federated networks on an intercontinental shared pool, while the trustworthiness monitoring and assessment platform will provide quantitative scores regarding the trustworthiness of an application running on the ATMOSPHERE ecosystem.

---

1 - ATMOSPHERE official website: www.atmosphere-eubrazil.eu

Primary authors: Dr ANTUNES, Nuno (University of Coimbra); BLANQUER, Ignacio; Dr BRASILEIRO, Francisco (Universidade Federal de Campina Grande); Dr CALATRAVA, Amanda (UPV); Dr VIEIRA, Marco (University of Coimbra)

Track Classification: R&D for computing services, networking, and data-driven science at the Iberian level.

Contribution ID: **8** Type: **Lightning talks**

A EOSC Demonstrator Project: Marine Eukaryote Genomics Portal

We will undertake the implementation of a genome annotation platform that will provide community access to tools and data-flows for marine genome annotation. The platform is designed to address the fragmented research landscape for genome annotation of marine organisms. We propose a portal to marine genomic resources and a community driven annotation platform for marine eukaryotes which would provide a focus for post-assembly genomic workflows and data access and complement access services such as EMBRIC Configurator, ELIXIR ontologies, and meta-data standards. Together these resources would expose the workflow from genome data collection to publication using open access and FAIR compliant standards and procedures. Although taxon agnostic, initially the platform will focus on pelagic fishes (the closely related Sardinha and Alosa) and the use primarily of comparative methods of gene prediction and validation.

Primary authors: Dr COX, Cymon (CCMAR); Dr LOURO, Bruno (CCMAR); Dr DE MORO, Gianluca (CCMAR); Prof. CANARIO, Adelino (CCMAR)

Track Classification: Development of applications for supporting User Communities in the context of the EOSC.

Contribution ID: **9** Type: **Lightning talks**

Design and deployment of a self-managed infrastructure for large-scale medical image analisys

Large-scale analysis of medical images using biomarkers requires an infrastructure which commonly exceeds the resources available for research groups. Besides, some biomarkers can benefit from specific hardware accelerators. Additionally, medical data analysis may require using only certified environments in specific countries, due to legal constraints. Cloud platforms enable medical institutions to use, paying by utilisation, several services like powerful machines, specific hardware and the guarantee of the execution in certified environments. This work describes the designed architecture for large-scale medical images analysis using biomarkers in Cloud platforms. Docker containers provide the developers with a way to encapsulate and deliver their applications and its dependencies for convenient distribution, so the biomarkers are encapsulated into Docker containers. The architecture involves all process of biomarker distribution pipeline: from updating the biomarker in the code repository, building the Docker image of the biomarker and executing it on a Cloud infrastructure. This infrastructure includes dynamic horizontal elasticity according to the jobs queue. Moreover, the infrastructure uses a large-scale distributed storage for accessing the data to be analysed.

Primary authors: LÓPEZ HUGUET, Sergio (Universitat Politècnica de València); BLANQUER, Ignacio; Dr ALBERICH-BAYARRI, Ángel (Quibim)

Track Classification: Development of applications for supporting User Communities in the context of the EOSC.

Contribution ID: **10** Type: **Presentation**

From AENEAS to the SKA Regional Centres. Spanish contribution

SKA is an international project, qualified as ESFRI Landmark Project, to build the largest and most sensitive radio telescope ever conceived, with the potential to achieve fundamental advances in Astrophysics, Physics and Astrobiology. Since 2011 the IAA-CSIC coordinates the Spanish participation in the SKA, closely collaborating with Portugal in SKA related activities during this time. Spain has recently become the eleventh Member of the SKA Organisation thus ensuring the participation of Spanish groups in the scientific exploitation of SKA data and in the construction of the telescope.

The SKA will also be the greatest data research public project, once complete. It will be composed of thousands of antennas distributed over distances of up to 3000 km, on both Africa and Australia and it will generate a copious data flux (around 1TB/s) that will turn the task of extracting scientifically relevant information into a scientific and technological Big Data challenge. The SKA Science Data Processor (SDP), will transform this flux of raw data into calibrated data products that will be delivered, at an average rate of about 150PB/year, to worldwide distributed data centres –called SKA Regional Centres (SRCs)- that not only will provide access to the SKA data but also to the analysis tools and processing power. The SRCs will have hence a key role in the exploitation of SKA data and the achievement of the SKA scientific goals.

An Alliance of SKA Regional Centres (SRCs) is being designed to address the challenge of scientifically exploiting the SKA data deluge. IAA-CSIC participates in different initiatives addressing this task, highlighting AENEAS, an H2020 on-going project. Its objective is to design a distributed and federated European SRC considering the existing services offered by European e-Infrastructures. AENEAS consortium includes a Portuguese partner, the Instituto de Telecomunicaçoes, thus being a framework to promote the Iberian collaboration in the context of the SRCs.

In addition, IAA-CSIC coordinates SKA-Link, a project that complements AENEAS efforts by studying how SRCs will face the challenge of supporting Open and Reproducible Science.

In this talk, we will present the contribution of IAA-CSIC to the AENEAS project and other activities related to the SRCs, including SKA-Link as well as our work studying how Distributed Computing Infrastructures (Ibercloud, EGI Federated Cloud and Amazon Web Services among others) fulfil requirements of a pipeline for calibrating data from LOFAR, one of the SKA pathfinders.

Primary authors: SÁNCHEZ EXPÓSITO, Susana (IAA-CSIC); Prof. VERDES-MONTENEGRO , Lourdes (IAA-CSIC); Dr GARRIDO SÁNCHEZ, Julian (IAA-CSIC)

Track Classification: Cooperation between Research Communities at the Iberian level

Contribution ID: **11** Type: **Presentation**

From AENEAS to SKA Regional Centres. Portuguese contribution

The Square Kilometre Array (SKA) project approaches construction for it Phase1. As part of science delivery, SKA1 through the AENEAS H2020 project aims the design and constitution of its EU Regional Centre with node in several SKA1 Members countries. Portuguese ENGAGE SKA RI partcpates in the design and specification of a distributed, European Science Data Centre (ESDC) to support the pan-European astronomical community in achieving the scientific goals of the SKA. AENEAS consortium includes a Spanish partner, the IAA-CSIC, thus being a framework to promote the Iberian collaboration in the context of the SRCs. We intend to highlight current contributions from Portugal and Ibearian collaboration with IAA-CSIC to this endeavor.

Primary authors: Mr BARBOSA, Domingos (Instiututo de Telecomunicações); Mr BARRACA, João Paulo (Instituto de telecomunicações); Mr MORGADO, Bruno (FCUP); Mr MAIA, Dalmiro (FCUP); Mr GOMES, Diogo (U. Aveiro); Mr BERGANO, Miguel (Instituto de Telecomunicações)

Track Classification: Cooperation between Research Communities at the Iberian level

Contribution ID: **12** Type: **Presentation**

Implementing High-Throughput Sequencing in bacterial foodborne pathogen surveillance: The INNUENDO Platform

Background: Outbreak investigations and pathogen surveillance are crucial tasks to control transmission of foodborne transmitted diseases. The decreasing costs of High-Throughput Sequencing (HTS) are boosting application of HTS for molecular typing in routine surveillance and outbreak investigation, maximizing discriminatory power in outbreak detection. However, lack of standardized bioinformatics infrastructures for data processing and integration, together with limited bioinformatics skills, continue to be major hurdles of HTS routine implementation, specially when analysing large datasets where the required computational needs are not available to most of the groups. To overcome these limitations, we developed the INNUENDO platform, an infrastructure that provides a user-friendly interface and the required framework for data analysis, from raw data quality assurance to integration of epidemiological data and visualization of the final analyses, providing the tools for the use of HTS techniques in everyday surveillance and outbreak investigation.
Methods: The INNUENDO platform is composed of two main applications that interact with each other using a REST API. The first application provides a graphical user interface that allows the user to define their own projects (groups of sequencing data and their associated metadata), control which protocols (software and their parameters) are applied to the data, and visualize the results. The second application controls the job submissions and status by using Nextflow as workflow engine and SLURM or any other job scheduler supported by Nextflow to control the available resources. Each software is available as a docker image and are loaded if required depending on the submitted job. The INNUENDO Platform includes the INNUca pipeline for automatic QC from reads to draft genome assemblies, which ultimately aims at producing consistently high-quality and comparable genomic data. The curated genome assemblies are then analysed following a gene-by-gene typing based approach. The chewBBACA software is used to perform the allele calling for whole genome MLST (wgMLST) profile definition. The wgMLST profiles generated for each isolate of interest can then be compared with profiles already stored in the platform's database. The wgMLST profiles of the isolates of interest, together with a selection of the closest ones in the database, are then filtered to produce a core genome MLST (cgMLST) and the data sent to PHYLOViZ Online for the construction of a minimum spanning tree annotated with metadata, allowing the exploration of possible epidemiological scenarios.
Results and conclusion: INNUENDO platform was developed with a modular design allowing the incorporation of different bioinformatic tools for the characterization of specific pathogens, and the capacity of being run in High Performance Computers clusters, which can greatly reduce the analysis time for large datasets. The modular nature of the platform implementation also allows for scalability in terms of computing needs. It also aims to facilitate data sharing and communication between different institutions, promoting cooperation in surveillance and outbreak investigation. The use of open source tools and standardized protocols will allow a future accreditation of the INNUENDO platform.
References and acknowledgements: More information on INNUENDO project (co-funded by EFSA) and the platform can be found in http://www.innuendoweb.org/. A working prototype of the INNUENDO platform was produced with the support of INCD funded by FCT and FEDER under the project 22153-01/SAICT/2016".

Primary authors: RIBEIRO-GONÇALVES, Bruno (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal); SILVA, Diogo (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal); P. MACHADO, Miguel (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal); SILVA, Mickael (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal); HALKILAHTI, Jani (National Institute for Health and Welfare, Helsinki, Finland); JAAKKONEN, Anniina (Microbiology Research Unit, Finnish Food Safety Authority, Evira, Finland); RAMIREZ, Mario (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portuga); ROSSI, Mirko (Department of Food Hygiene and Environmental Health, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland); ANDRÉ CARRIÇO, João (Instituto de Microbiologia and Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal)

Track Classification: R&D for computing services, networking, and data-driven science at the Iberian level.

Contribution ID: **13** Type: **Lightning talks**

A New Role for Supercomputing Centers in Open Science

Thursday 11 October 2018 17:00 (10 minutes)

In the last decades an exponential increase of the scientific and technical development in all the areas of science has become manifest, being more and more relevant the conflicts between the pure scientific advance of the society and the property of the researched knowledge. Particularly the brake that certain aspects of the established system suppose in the acquisition of some findings. The ever demanding scientific community aims for new platforms and services up to date in order to diminish the inconveniences that have arisen due this same growth. The guidelines provided by the Open Science (OS), despite being a concept already discuss for some time, set out the perfect framework for the creation of new applications and platforms in which the research centers might take over from the responsibility exercised so far by the publishers to allocate resources for disclosure and distribution without replace them. In particular Open Access (OA), Open Data (OD) and Open Methodologies (OM) are the guidelines that can fit the most in the current tasks performed by supercomputing and research centers where availability, reliability and security are concepts well implemented already and can be very powerful skills to them in order to step in the spotlight in this new framework which is OS. In the other hand, while universities and publishers can be tempted by recognition and self promotion at the time of deciding whether a job can benefit them or not, the public supercomputing centers, that are already providing services for both private and public projects, working in the same direction and as a set, could be considered more objectives. In this paper it is proposed a path to follow by the supercomputing centers, within the framework of European Open Science Cloud (EOSC), to share resources with the aim of providing an adequate infrastructure for the development of scientific research, adding to their current competences the ability to become neutral ground for scientific disclosure.

Presenter: JIMÉNEZ, Luis Ignacio

Session Classification: Lightning talks

Contribution ID: **14** Type: **Lightning talks**

Design and deployment of a self-managed infrastructure for large-scale medical image analisys

Thursday 11 October 2018 17:10 (10 minutes)

Large-scale analysis of medical images using biomarkers requires an infrastructure which commonly exceeds the resources available for research groups. Besides, some biomarkers can benefit from specific hardware accelerators. Additionally, medical data analysis may require using only certified environments in specific countries, due to legal constraints. Cloud platforms enable medical institutions to use, paying by utilisation, several services like powerful machines, specific hardware and the guarantee of the execution in certified environments. This work describes the designed architecture for large-scale medical images analysis using biomarkers in Cloud platforms. Docker containers provide the developers with a way to encapsulate and deliver their applications and its dependencies for convenient distribution, so the biomarkers are encapsulated into Docker containers. The architecture involves all process of biomarker distribution pipeline: from updating the biomarker in the code repository, building the Docker image of the biomarker and executing it on a Cloud infrastructure. This infrastructure includes dynamic horizontal elasticity according to the jobs queue. Moreover, the infrastructure uses a large-scale distributed storage for accessing the data to be analysed.

Presenter: LÓPEZ HUGUET, Sergio (Universitat Politècnica de València)

Session Classification: Lightning talks

Contribution ID: **15** Type: **not specified**

A EOSC Demonstrator Project: Marine Eukaryote Genomics Portal

Thursday 11 October 2018 17:20 (10 minutes)

We will undertake the implementation of a genome annotation platform that will provide community access to tools and data-flows for marine genome annotation. The platform is designed to address the fragmented research landscape for genome annotation of marine organisms. We propose a portal to marine genomic resources and a community driven annotation platform for marine eukaryotes which would provide a focus for post-assembly genomic workflows and data access and complement access services such as EMBRIC Configurator, ELIXIR ontologies, and meta-data standards. Together these resources would expose the workflow from genome data collection to publication using open access and FAIR compliant standards and procedures. Although taxon agnostic, initially the platform will focus on pelagic fishes (the closely related Sardinha and Alosa) and the use primarily of comparative methods of gene prediction and validation.

Presenter: COX, Cymon (CCMAR (Centro de Ciencias do Mar))

Session Classification: Lightning talks

Contribution ID: **16** Type: **not specified**

Scipion on-demand service in the cloud

Friday 12 October 2018 09:30 (20 minutes)

Scipion is an image processing framework used to obtain 3D maps of macromolecular complexes on Cryo Electron Microscopy. It has emerged as the solution offered by the Instruct Image Processing Center (I2PC), hosted by CNB-CSIC, to European scientists accessing the European Research Infrastructure for Structural Biology (Instruct).

Cryo-EM processing is very demanding in terms of computing resources requiring powerful servers and since recently the use of GPUs. Common desktop machines are clearly insufficient in computing capability and storage which could a problem for many scientists that might not have access to big servers or GPUs.

Cloud IaaS (Infrastructure as a Service) is a new form of accessing computing and storage resources on demand. To effectively use cloud infrastructures ScipionCloud was developed, resulting in a full installation of Scipion both in public and private clouds, accessible as public "images", that include all needed cryoEM software and just requires a Web browser to work as if it was a local desktop. These images are available in the EGI Applications Database and in AWS public AMIs catalogue.

We present here a new service for Instruct users that would allow them to process the data acquired at any of the high end Instruct Facilities -focusing this initial work in own cryo EM Facility at the I2PC- on a virtual machine in one of the IberGRID sites. In this first scenario we are now presenting, the machine itself is setup by I2PC staff, but as we advance in our development we envision the opening of a web portal accessible to I2PC users to do that.

Presenter: DEL CANO, Laura

Session Classification: Applications and Cooperative development

Contribution ID: **17** Type: **not specified**

From AENEAS to the SKA Regional Centres. Spanish contribution

Friday 12 October 2018 09:50 (20 minutes)

SKA is an international project, qualified as ESFRI Landmark Project, to build the largest and most sensitive radio telescope ever conceived, with the potential to achieve fundamental advances in Astrophysics, Physics and Astrobiology. Since 2011 the IAA-CSIC coordinates the Spanish participation in the SKA, closely collaborating with Portugal in SKA related activities during this time. Spain has recently become the eleventh Member of the SKA Organisation thus ensuring the participation of Spanish groups in the scientific exploitation of SKA data and in the construction of the telescope.

The SKA will also be the greatest data research public project, once complete. It will be composed of thousands of antennas distributed over distances of up to 3000 km, on both Africa and Australia and it will generate a copious data flux (around 1TB/s) that will turn the task of extracting scientifically relevant information into a scientific and technological Big Data challenge. The SKA Science Data Processor (SDP), will transform this flux of raw data into calibrated data products that will be delivered, at an average rate of about 150PB/year, to worldwide distributed data centres –called SKA Regional Centres (SRCs)- that not only will provide access to the SKA data but also to the analysis tools and processing power. The SRCs will have hence a key role in the exploitation of SKA data and the achievement of the SKA scientific goals.

An Alliance of SKA Regional Centres (SRCs) is being designed to address the challenge of scientifically exploiting the SKA data deluge. IAA-CSIC participates in different initiatives addressing this task, highlighting AENEAS, an H2020 on-going project. Its objective is to design a distributed and federated European SRC considering the existing services offered by European e-Infrastructures. AENEAS consortium includes a Portuguese partner, the Instituto de Telecomunicaçoes, thus being a framework to promote the Iberian collaboration in the context of the SRCs.

In addition, IAA-CSIC coordinates SKA-Link, a project that complements AENEAS efforts by studying how SRCs will face the challenge of supporting Open and Reproducible Science.

In this talk, we will present the contribution of IAA-CSIC to the AENEAS project and other activities related to the SRCs, including SKA-Link as well as our work studying how Distributed Computing Infrastructures (Ibercloud, EGI Federated Cloud and Amazon Web Services among others) fulfil requirements of a pipeline for calibrating data from LOFAR, one of the SKA pathfinders.

Presenter: SÁNCHEZ EXPÓSITO, Susana (IAA-CSIC)

Session Classification: Applications and Cooperative development

Contribution ID: **18** Type: **not specified**

From AENEAS to SKA Regional Centres. Portuguese contribution

The Square Kilometre Array (SKA) project approaches construction for it Phase1. As part of science delivery, SKA1 through the AENEAS H2020 project aims the design and constitution of its EU Regional Centre with node in several SKA1 Members countries. Portuguese ENGAGE SKA RI partcpates in the design and specification of a distributed, European Science Data Centre (ESDC) to support the pan-European astronomical community in achieving the scientific goals of the SKA. AENEAS consortium includes a Spanish partner, the IAA-CSIC, thus being a framework to promote the Iberian collaboration in the context of the SRCs. We intend to highlight current contributions from Portugal and Ibearian collaboration with IAA-CSIC to this endeavor.

Presenter: BARBOSA, Domingos

Session Classification: Applications and Cooperative development

Contribution ID: **19** Type: **not specified**

HPC and Cloud Resources for Running Mathematical Simulations Efficiently

Friday 12 October 2018 11:30 (20 minutes)

Complex simulations that require large amounts of computational resources have typically run in dedicated supercomputers. However some parts of these simulations don´t perform well in these computers or don´t need these highly costs resources and can be executed on cheaper hardware. Moving some parts of these simulations out of the supercomputers and running them in smaller clusters or cloud resources can improve the time to results and reduce the costs of the simulations, providing also higher flexibility and ease of usage. Both HPC and Cloud resources benefit and empower users who need to perform complex simulations that normally only take advantage of the capabilities of one of these infrastructures. We propose the combined usage of these platforms by using an orchestrator to coordinate the exploitation of these systems and container technology to enable interoperability between them. Such solution provides simulations as a service in a transparent way for end-users and software developers, as well as improves the efficiency in HPC resources usage. It has been proven to work with different HPC and Cloud providers, including EOSC Hub.

Presenter: CARNERO, Javier (Atos Research & Innovation)

Session Classification: R & D in Computing Centers

Contribution ID: **20** Type: **not specified**

e-Science services evolution at RedIRIS

Friday 12 October 2018 11:50 (20 minutes)

This presentation will outline the services currently in production, and also under development, oriented to support the access of researchers in Spain to the Digital Infrastructures of the EOSC era.The purpose of the presentation is also collecting feedback from the participants.

Presenter: FUENTES, Antonio

Session Classification: R & D in Computing Centers

Contribution ID: **21** Type: **not specified**

Easy use of Distributed TensorFlow Training on supercomputing facilities

Friday 12 October 2018 12:10 (20 minutes)

Deep Learning is a powerful tool for science, industry and other sectors that benefits from large datasets and computing capacity during models'design and training phases. TensorFlow (TF) Google's Machine Learning API is one of the tools most widely used for developing and training such deep learning models. There is a wide range of possibilities to configure a deep learning model however find the optimal model architecture can be a highly demanding computing task. Moreover, when involved datasets are very large the computing requirements increase and training processes can take a lot of time and hinder the design cycle. One of the most powerful capabilities of TF is its distributed computing capabilities, allowing portions of the automatic generated graph to be calculated on different computing nodes, and speeding up the training process. Deployment of distributed TF is not a straightforward task and it presents several issues, mainly related with its use under the control of local resources management systems and the usage of the right resources. In order to allow CESGA users to adapt their own TF codes to take advantage of TF and Finis Terrae II distributed computing capabilities, a complete python Toolkit has been developed. This Toolkit deals with several tasks that are not relevant in the models design, but necessary for exploiting the distributed capabilities, hiding the underlying complexity to final users. Additionally, an example of a successful industrial case, based on the Fortissimo 2 project experiment "Cyber-Physical Laser Metal Deposition (CyPLAM)", that uses this Toolkit, is presented. Thanks to the TF distributed capabilities, the computing capability of Finis Terrae and the use of the developed Toolkit, the time needed for training the largest model of this industrial case has been decreased from 8 hours (non- distributed TF) to less than 20 minutes.

Presenter: FERRO, Gonzalo (CESGA)

Session Classification: R & D in Computing Centers

Contribution ID: **22** Type: **not specified**

To Trust or Not To Trust, that is the question

Friday 12 October 2018 12:30 (20 minutes)

Trust is a central issue confronting men and women in contemporary society. In fact, the most difficult thing to achieve in this world is trust. It can take years to win and only a matter of seconds to lose it. This is also applicable in a computing environment, where users need to trust computing services to process and manage their data. This implies a broad spectrum of properties to be accomplished, such as Security, Privacy, Coherence, Isolation, Stability, Fairness, Transparency and Dependability.
Adaptive, Trustworthy, Manageable, Orchestrated, Secure Privacy-assuring Hybrid, Ecosystem for REsilient Cloud Computing1 (hereinafter "ATMOSPHERE") is a European-Brazilian collaboration project aiming at measuring and improving the different trustworthiness dimensions of data analytics applications running on the cloud. To achieve trustworthy cloud computing services on a federated environment, ATMOSPHERE focuses on providing four components: i) a dynamically reconfigurable hybrid federated VM and container platform, to provide isolation, high-availability, Quality of Service (QoS) and flexibility; ii) Trustworthy Distributed Data Management services that maximise privacy when accessing and processing sensitive data; iii) Trustworthy Distributed Data Processing services to build up and deploy adaptive applications for Data Analytics, providing high-level trustworthiness metrics for computing fairness and explainability properties; and iv) a Trustworthy monitoring and assessment platform, to compute trustworthiness measures from the metrics provided by the different layers.
In this lightning session, we will focus our discussion in the integration of the federated cloud platform with the Trustworthy monitoring and assessment platform, in order to provide isolation, stability and Quality of Service performance guarantees. The cloud platform will enable the dynamic reconfiguration of resource allocation to applications running on federated networks on an intercontinental shared pool, while the trustworthiness monitoring and assessment platform will provide quantitative scores regarding the trustworthiness of an application running on the ATMOSPHERE ecosystem.

1 - ATMOSPHERE official website: www.atmosphere-eubrazil.eu

Presenter: Dr CALATRAVA, Amanda (UPV)

Session Classification: R & D in Computing Centers

Contribution ID: **23** Type: **not specified**

Implementing High-Throughput Sequencing in bacterial foodborne pathogen surveillance: The INNUENDO Platform

Background: Outbreak investigations and pathogen surveillance are crucial tasks to control transmission of foodborne transmitted diseases. The decreasing costs of High-Throughput Sequencing (HTS) are boosting application of HTS for molecular typing in routine surveillance and outbreak investigation, maximizing discriminatory power in outbreak detection. However, lack of standardized bioinformatics infrastructures for data processing and integration, together with limited bioinformatics skills, continue to be major hurdles of HTS routine implementation, specially when analysing large datasets where the required computational needs are not available to most of the groups. To overcome these limitations, we developed the INNUENDO platform, an infrastructure that provides a user-friendly interface and the required framework for data analysis, from raw data quality assurance to integration of epidemiological data and visualization of the final analyses, providing the tools for the use of HTS techniques in everyday surveillance and outbreak investigation.
Methods: The INNUENDO platform is composed of two main applications that interact with each other using a REST API. The first application provides a graphical user interface that allows the user to define their own projects (groups of sequencing data and their associated metadata), control which protocols (software and their parameters) are applied to the data, and visualize the results. The second application controls the job submissions and status by using Nextflow as workflow engine and SLURM or any other job scheduler supported by Nextflow to control the available resources. Each software is available as a docker image and are loaded if required depending on the submitted job. The INNUENDO Platform includes the INNUca pipeline for automatic QC from reads to draft genome assemblies, which ultimately aims at producing consistently high-quality and comparable genomic data. The curated genome assemblies are then analysed following a gene-by-gene typing based approach. The chewBBACA software is used to perform the allele calling for whole genome MLST (wgMLST) profile definition. The wgMLST profiles generated for each isolate of interest can then be compared with profiles already stored in the platform's database. The wgMLST profiles of the isolates of interest, together with a selection of the closest ones in the database, are then filtered to produce a core genome MLST (cgMLST) and the data sent to PHYLOViZ Online for the construction of a minimum spanning tree annotated with metadata, allowing the exploration of possible epidemiological scenarios.
Results and conclusion: INNUENDO platform was developed with a modular design allowing the incorporation of different bioinformatic tools for the characterization of specific pathogens, and the capacity of being run in High Performance Computers clusters, which can greatly reduce the analysis time for large datasets. The modular nature of the platform implementation also allows for scalability in terms of computing needs. It also aims to facilitate data sharing and communication between different institutions, promoting cooperation in surveillance and outbreak investigation. The use of open source tools and standardized protocols will allow a future accreditation of the INNUENDO platform.
References and acknowledgements: More information on INNUENDO project (co-funded by EFSA) and the platform can be found in http://www.innuendoweb.org/. A working prototype of the INNUENDO platform was produced with the support of INCD funded by FCT and FEDER under the project 22153-01/SAICT/2016".

Presenter: GONÇALVES, Bruno (Instituto de Medicina Molecular)

Session Classification: R & D in Computing Centers

Contribution ID: **24** Type: **Presentation**

IBERLIFE-IBERGRID initiative

Thursday 11 October 2018 15:30 (30 minutes)

Presenter: Dr GONZALEZ-ARANDA, Juan Miguel (Lifewatch ERIC)

Session Classification: IBERGRID 2018 Welcome Plenary

Contribution ID: **25** Type: **Presentation**

High-resolution coastal modeling and forecasting using HPC: lessons learned from a decadal experience

Friday 12 October 2018 09:00 (30 minutes)

Estuaries and coastal zones are among the most productive ecosystems on Earth, supporting many human activities and providing multiple ecosystem services. The ability to simulate and forecast the dynamics of estuarine and coastal zones is thus essential to support the sustainable management of these regions, both for daily activities and for long-term strategies associated with climate change.
Computational forecast systems are an important asset to address these concerns by providing predictions of relevant variables, through the integration of numerical models and field data. The reliability of the forecast predictions depends however on the accuracy of the models behind them. Unstructured grid numerical models have been used for several decades to simulate coastal zones at LNEC to address the need for adequate spatial and temporal discretizations. For the past decade these models have been integrated in LNEC's forecast platform WIFF (WIFF - Water Information Forecast Framework) to predict water circulation and water quality in coastal zones, taking advantage of the resources of the Portuguese National Computational Infrastructure (INCD).
This communication summarizes and evaluates our experience of running forecast systems in the INCD. The applications range from the inundation of estuarine margins to oil spill and water contamination predictions. End-users include the Civil Protection agency, port authorities and wastewater utilities. The major focus will be performance issues (comparing grid and cloud resources), service level performance and user experience.

Primary authors: Dr OLIVEIRA, Anabela (LNEC); ROGEIRO, Joao (LNEC); AZEVEDO, Alberto (LNEC); FORTUNATO, Andre; RODRIGUES, Marta; TEIXERIA, Joana; GOMES, Jorge; DAVID, Mario; PINA, Joao; MARTINS, Joao Paulo

Session Classification: Applications and Cooperative development

Track Classification: Development of applications for supporting User Communities in the context of the EOSC.

Contribution ID: **26** Type: **Presentation**

IBERGRID: A personal retrospective

Thursday 11 October 2018 14:30 (30 minutes)

Presenter: GARCIA TOBIO, Javier

Session Classification: IBERGRID 2018 Welcome Plenary

Contribution ID: **27** Type: **not specified**

Ibergrid Status Report

Thursday 11 October 2018 15:00 (30 minutes)

Presenter: GOMES, Jorge (LIP)

Session Classification: IBERGRID 2018 Welcome Plenary

Contribution ID: **28** Type: **not specified**

IBERGRID towards EOSC

Presenter: CAMPOS, Isabel (CSIC)

Session Classification: IBERGRID 2018 Welcome Plenary

Contribution ID: **29** Type: **Presentation**

On-premises Serverless Container-aware ARchitectures

Friday 12 October 2018 12:50 (20 minutes)

Presenter: PÉREZ, Alfonso

Session Classification: R & D in Computing Centers