



LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

LIP Computing & Big Data

Joao Pina

LIP Distributed Computing and Digital Infrastructures group



WHAT

LIP IS

The reference institution for experimental particle physics in Portugal and the Portuguese reference partner of CERN as well as other international scientific infrastructures

Has centres in:
Lisbon, Coimbra, Braga



Experimental particle and astroparticle physics



Development of new instruments and methods



Scientific computing



Knowledge transfer, education and outreach

To whom we are connected



SNOLAB, SURF, MIT,
Queen's, Fermilab,
UBrown, UFlorida,
URockefeller, Auger,
CBPF, SBF, USP,
USC, UCampinas,
EERJ

UTaiwan,
IIT Madras

IST, FCUL, ULisboa, FCTUC, UM,
CTN, UA, ICNAS, LNEC, Ciência
Viva, IBEB, INESC-ID,
INESC-TECH, UBI, UÉvora, SPF,
ISEC/IPC, LIBPhys, BioSI,
CCMAR, ISEC, UPorto, IMM, IGC,
PORBIODATA, FCT-FCCN

CERN, ESA, EGI, DESY, HIP
Helsinki, MEPH, Imperial
College, USurrey, UOxford,
TUDresden, LMU Munich,
HepHYVienna, TUDortmund,
IPPP, LPC, TUDelft, GSI,
Humboldt, KIT, CEA,
CESNET,
Clermont-Ferrand,
CYFRONET, PSNC,
Utrecht



CSIC, IFCA, UPV, CESGA, BIFI, UAM, PIC, Lifewatch ESFRI,
UGranada, USC/IGFAE, INFN, INAF, UFerrara, UTorino,
UPadova, UPisa, UUdine, PoliMilano, PoliBari, LLR

Delivering IT services

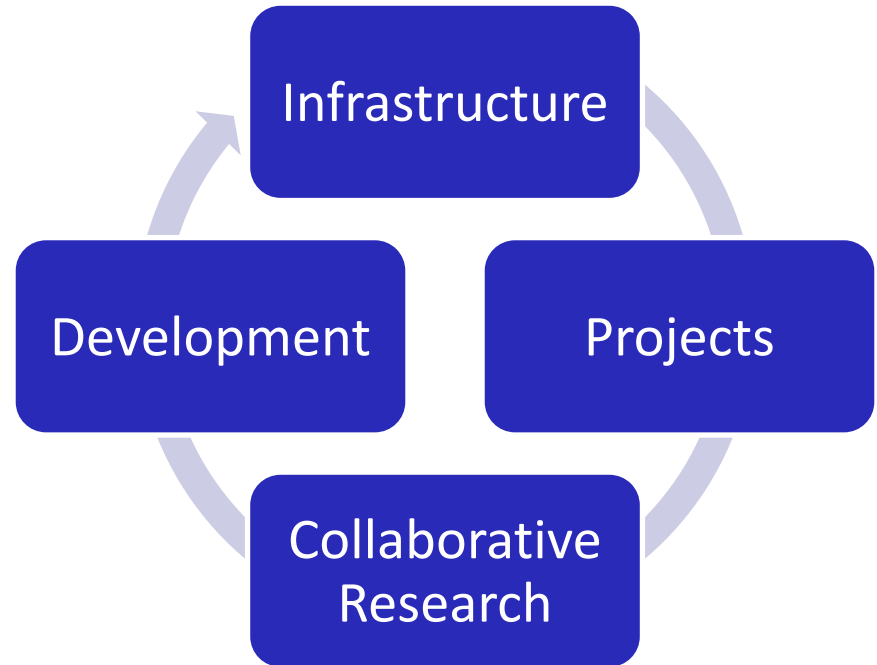
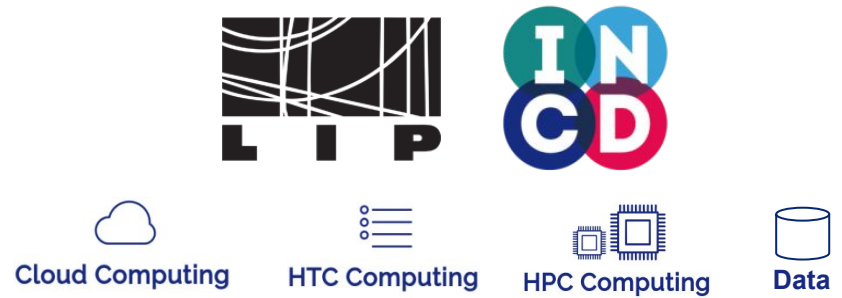
Computing and data

For LIP

1. **Delivering internal IT services to LIP**
 - Supporting research, innovation, education, outreach and administrative activities at LIP.
2. **Participation in national and international projects, initiatives and digital infrastructures**
 - Performing Research and Innovation

Via INCD

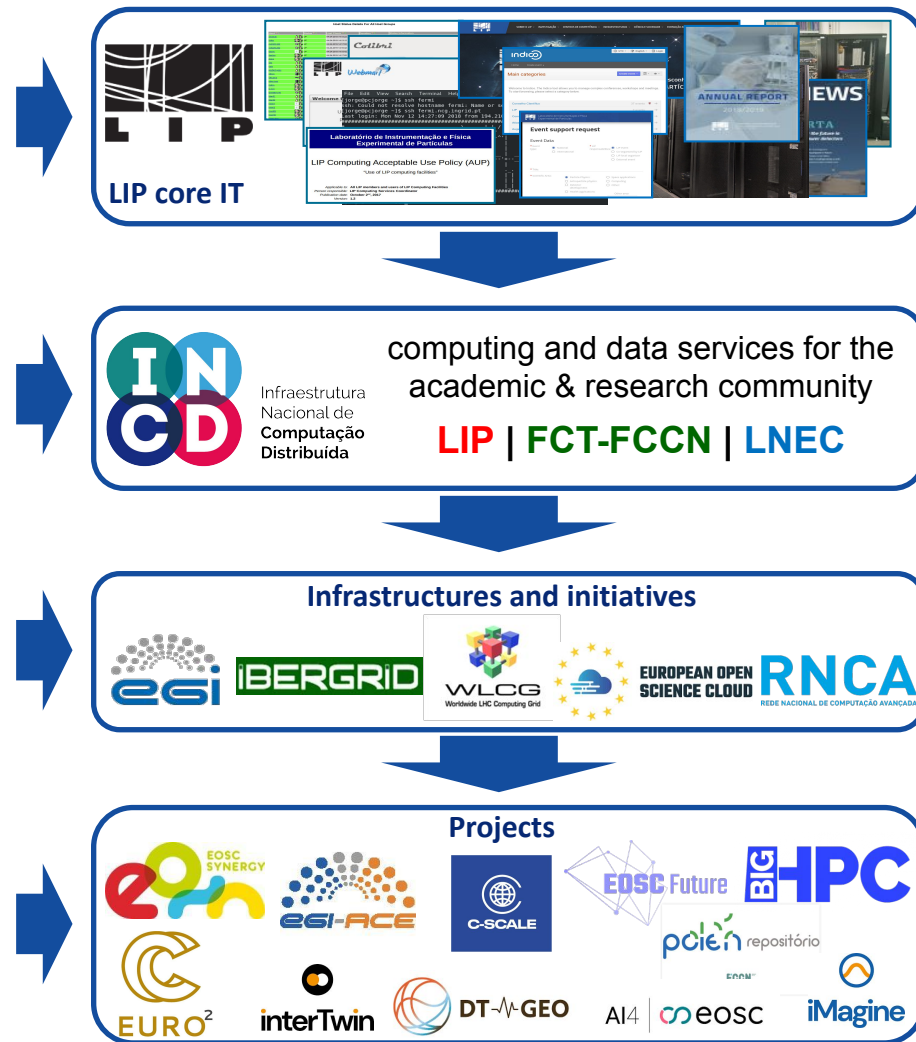
1. **Provide compute and data oriented services to Portuguese academic and research community**
 - INCD is a spinoff from the LIP participation in projects on computing and digital infrastructures.
 - Is the basis for the LIP computing services.
 - INCD itself is a legal entity, private non-profit association that joins LIP, FCT and LNEC.
 - Research infrastructure in the FCT roadmap of research infrastructures (until 2022).



Competences Projects & Initiatives

The group has participation in projects, infrastructures and initiatives in areas such as:

- Digital infrastructures
- High Performance Computing
- High Throughput Computing
- Cloud Computing
- Federated/distributed computing
- Software Quality Assurance
- Software management
- Linux containers
- Data repositories





INCD operations centers in 2024



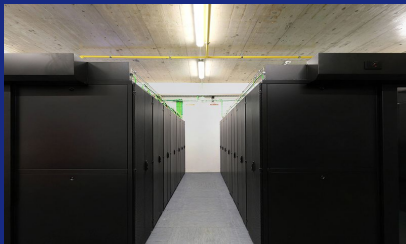
INCD-A @ LNEC in Lisbon
HPC / HTC / Cloud / Federation
6000 CPU cores
5 Petabytes online raw
100 Gbps
Includes the WLCG Tier-2



INCD-B @ REN in Riba-de-Ave
(DECOMMISSIONED)
HPC / HTC
2600 CPU cores
384 Terabytes raw
1 Gbps



INCD-L @ LIP in Lisbon
Tape storage
1 Petabyte backups
10 Gbps



INCD-D @ UTAD in Vila Real
(STARTED PROD 2024)
HPC / HTC / Cloud / Federation
5000 CPU cores + IB HDR200
4 Petabytes online raw
10 Gbps

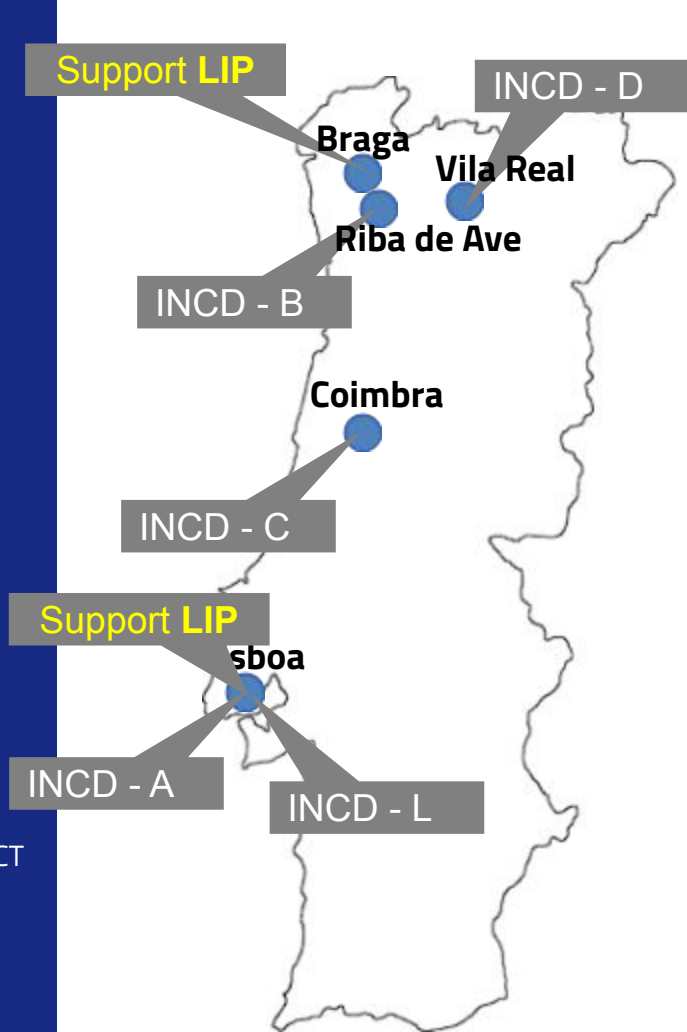


INCD-C @ UC in Coimbra
(BEING RENEWED)
Tape storage expansion
20 Petabytes
10 Gbps



Network and housing in 2024

- INCD - A @ “Sala-Grid” LNEC in Lisbon
 - Main data centre for INCD
 - 100 Gbps connectivity
 - Data centre managed by FCCN
 - Equipment managed by LIP team in Lisbon
- ~~INCD - B @ NOS/REN/EDP in Riba de Ave~~
 - ~~Bob supercomputer offered by TACC~~
 - ~~1 Gbps connectivity~~
 - ~~Commercial data centre under contract with FCT~~
 - ~~Equipment managed by LIP/INCD team in Braga/Minho~~
- INCD - C @ UC in Coimbra
 - 10 Gbps connectivity
 - Physics department datacenter under agreement with LIP
 - Equipment managed by LIP team in Lisbon
- INCD - D @ UTAD in Vila Real
 - 10 Gbps connectivity
 - University of Trás-os-Montes e Alto Douro data centre under agreement with FCT
 - Equipment managed by LIP/INCD team in Braga/minho
- INCD - L @ LIP in Lisbon
 - 10 Gbps connectivity
 - LIP datacenter
 - Equipment managed by LIP team in Lisbon

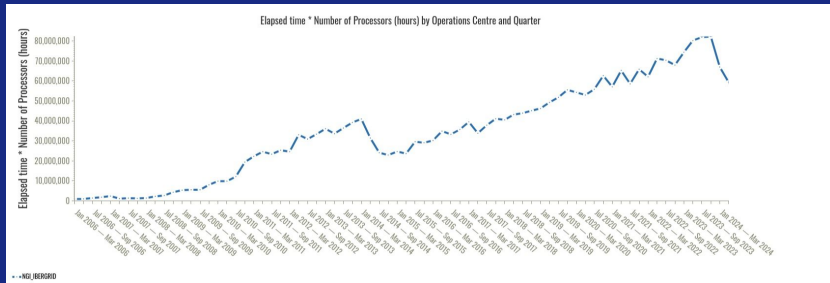




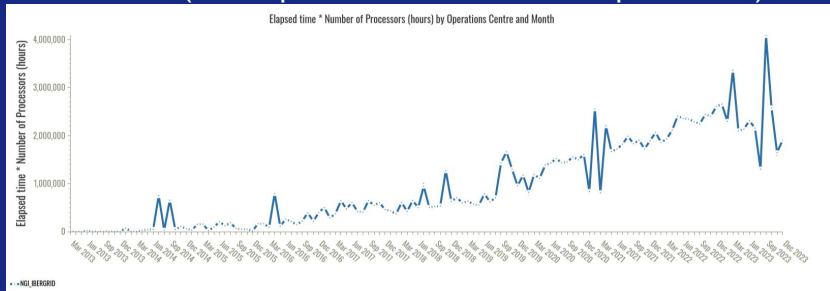
Federation in IBERGRID and EGI

INCD is the main Portuguese infrastructure in EGI and IBERGRID

IBERGRID - 320 million jobs since 2006 (INCD supported > 6%)



IBERGRID - 1.3 million instantiated VMs since 2015 (INCD provided 24% of the elapsed time)



Federating about 1,243,400 CPU cores and over 1,4 Exabyte of storage space from hundreds of data centres including WLCG

36 EU funded projects

1790¹ Enabled articles

7.1B HTC CPU hours consumed

29 EGI Council participants

84.000 Users

407M Computational jobs

70M Cloud CPU hours consumed

cloud+grid+data



IBERGRID and EGI

IBERGRID responsibilities and activities:

- IBERGRID and EGI provide the backbone for WLCG
- Operations coordination at Iberian level and interface with EGI operations
- Software management for the EGI and IBERGRID federations
- Regional technical contact point
- Support for user communities
- Developing and operating core services e.g. software repositories for the EGI federation
- Integration of thematic and/or user services

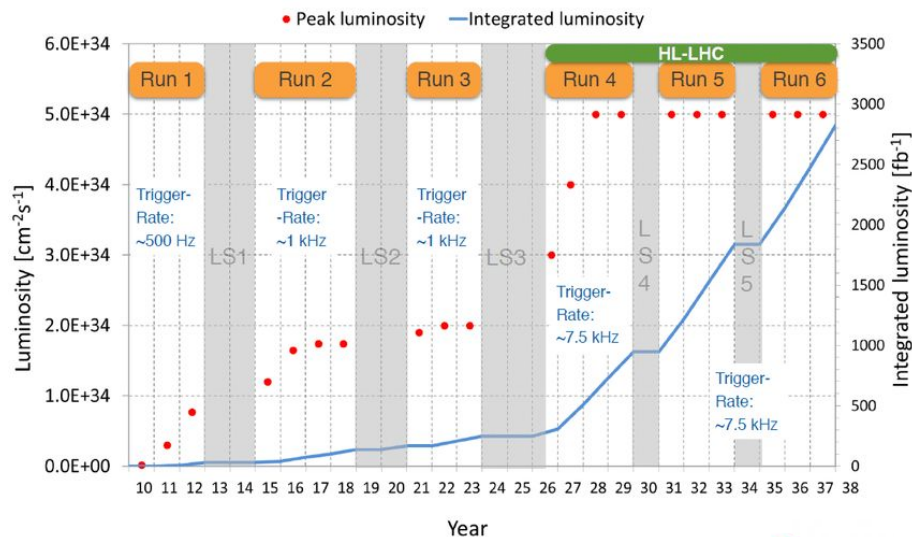


The challenge of Computing for the LHC

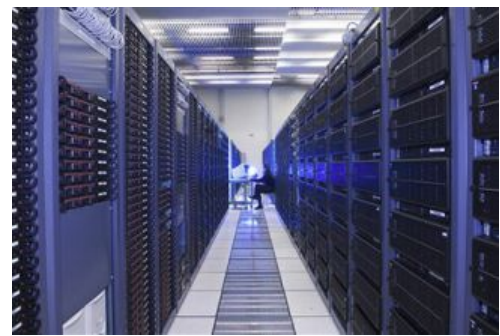
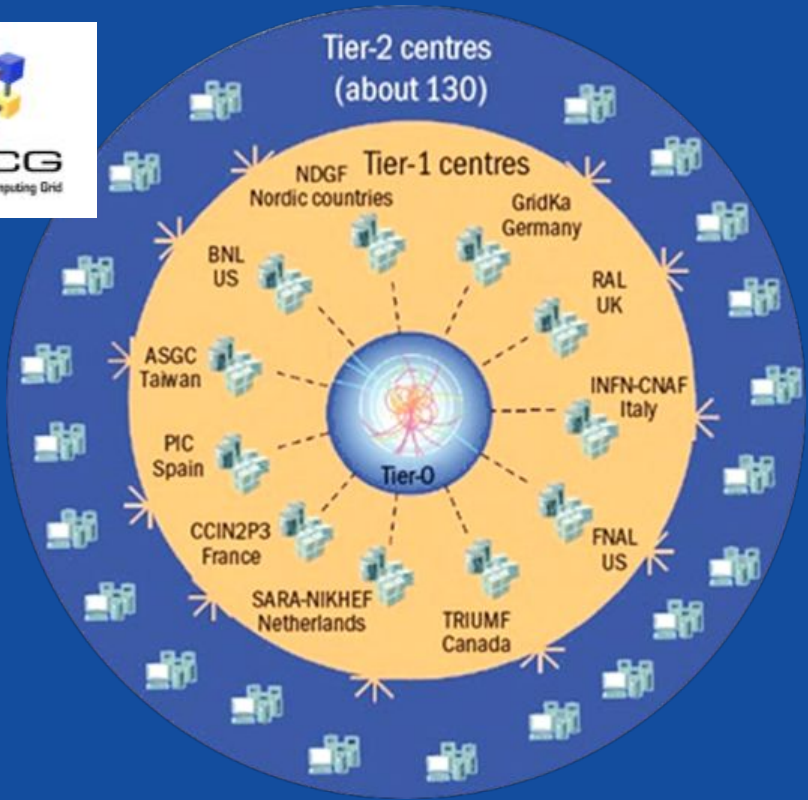
LIP participates in the **ATLAS** and **CMS** experiments at the CERN Large Hadron Collider (LHC).

The MoU for Collaboration in the Deployment and Exploitation of the WLCG was signed in 2006 by CERN, GRICES and LIP.

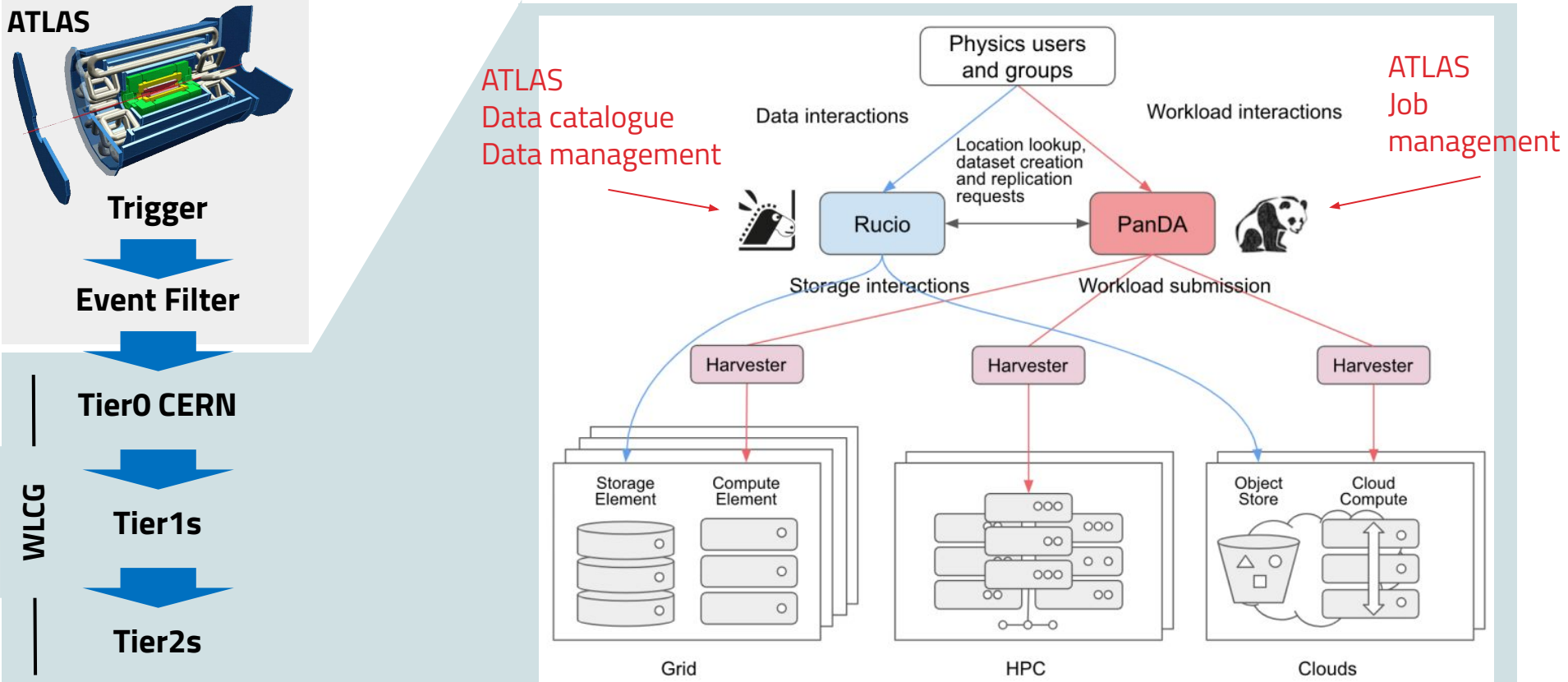
Under the MoU LIP operates a Tier-2 computing and data facility integrated in WLCG for ATLAS and CMS.



WLCG architecture



Overview of the data processing in the ATLAS experiment



Portuguese Tier-2 ATLAS and CMS

The Tier-2 uses grid middleware and operates in the framework of the INCD infrastructure.

Currently based on the INCD Slurm cluster in Lisbon (INCD-A @ LNEC).

- ARC-CE + Slurm batch scheduler
- XRootd + Lustre
- StoRM + Lustre
- CVMFS

ARC-CE



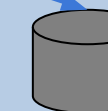
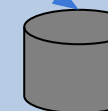
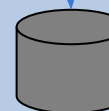
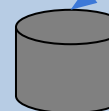
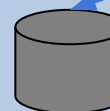
Slurm Scheduler



Slurm Farm



Xrootd
Webdav
SRM
etc

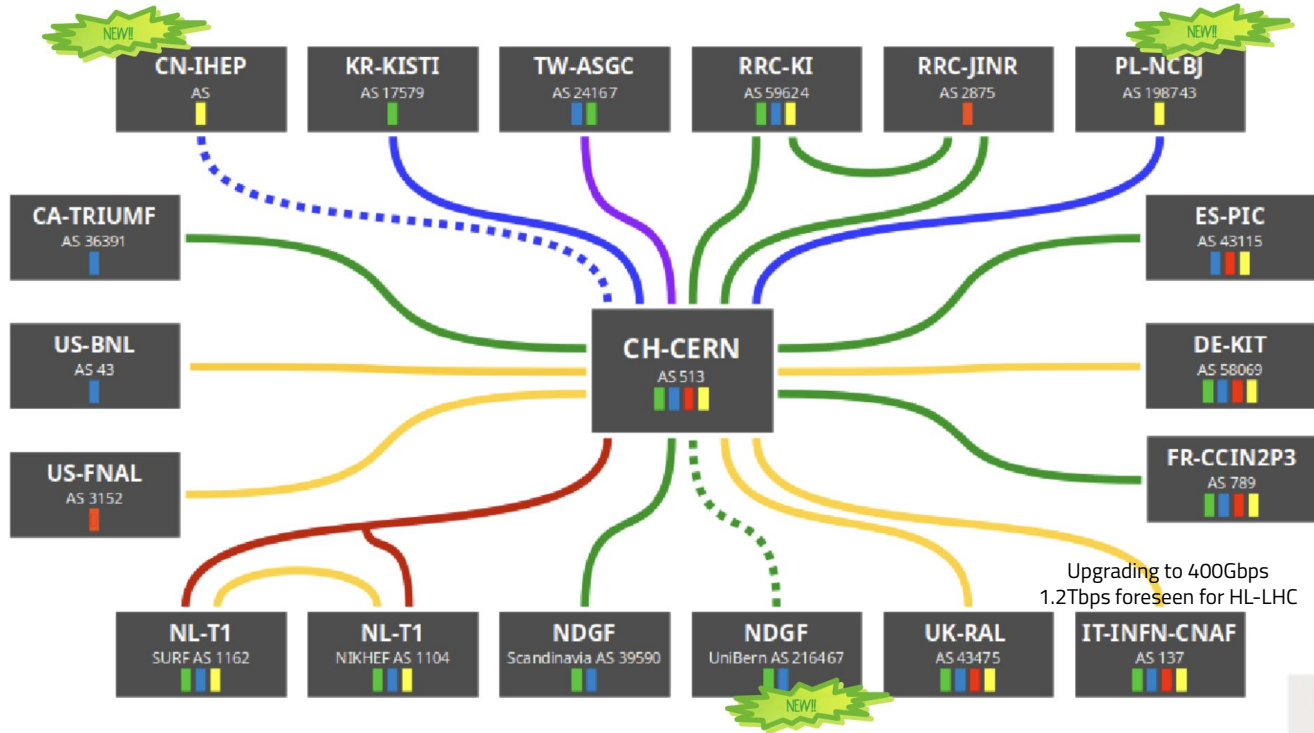


Lustre Data storage



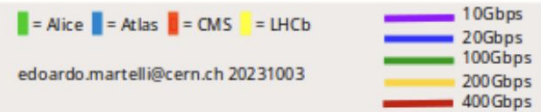
LHCOPN

LHC Optical Private Network (LHCONE)
Interconnects the WLCG Tier-1 centres worldwide



Star topology

- 1x Tier-0
- 15x Tier-1s
- 18x Sites
- 15x countries
- 3x continents



Requirement for 2029: 1 Tbps per site

LHCONE

LHC Open Network Environment (L3VPN service)

Private Network connecting WLCG sites (Tier-1s + Tier-2s)

- Dedicated network for LHC
 - **ATLAS, CMS**, ALICE LHCb
- A collaborative effort among Research & Education Network Providers
- Multi domain **L3 VPN**
- Routed Virtual Private Network connecting Science-DMZs
- BGP communities for traffic engineering
- Only for declared IP prefixes

The LHCONE is open to other HEP scientific collaborations

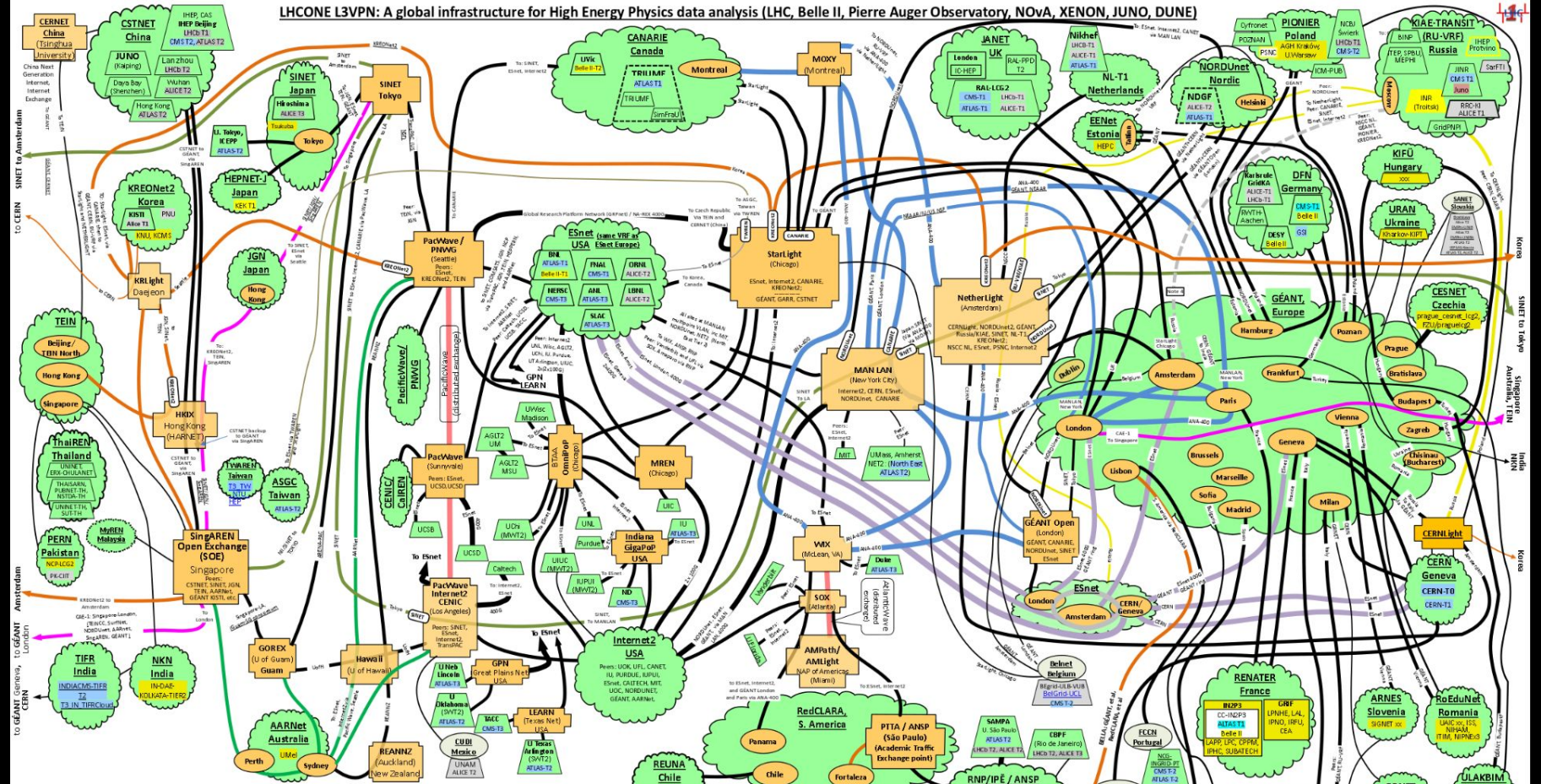
- Belle II
- Xenon
- Juno
- Nova
- **Pierre Auger Observatory**
- **DUNE**

The

- > 33 countries
- > 144 sites



LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOVA, XENON, JUNO, DUNE)



LHCONE Map Ver. 9.0, 2024-03-24 - WEJohnston, ESnet, wej@es.net

Legend:

- Green circle:** LHCONE VRF domain/aggregator
- Orange circle:** Provider network
- Blue circle:** Connector network or institution - provides, e.g., an I2 path between VRFs.
- Yellow circle:** Provider network PoP router
- Red circle:** WLCG sites that are not connected to LHCONE
- Black circle:** Exchange point
- Grey circle:** Future site

Line styles:

- Solid line:** Various links
- Dashed line:** Underlined link information indicates link provider, not use
- Double dash line:** Double dash outline indicates distributed site
- Thin solid line:** Future site

International infrastructure by provider/collaboration

ANASNET	SINET
GEANT	NORDUnet
SINET, Japan, global ring	KfAE, Russia
ESnet transatlantic, USA	KREONET2, Korea
SINET/JEN/SingAREN	BELLA: GEANT, Korea, RedCLARA, et al.

Legend for LHCONE VRFs:

- Blue:** LHC ALICE or LHCB site
- Green:** LHC Tier 1 ATLAS and CMS
- Orange:** LHC Tier 2/3 ATLAS and CMS
- Yellow:** Belle II Tier 1/2
- Red:** JUNO

Legend for sites that are standalone VRFs:

- Green:** JLAB

NOTES:

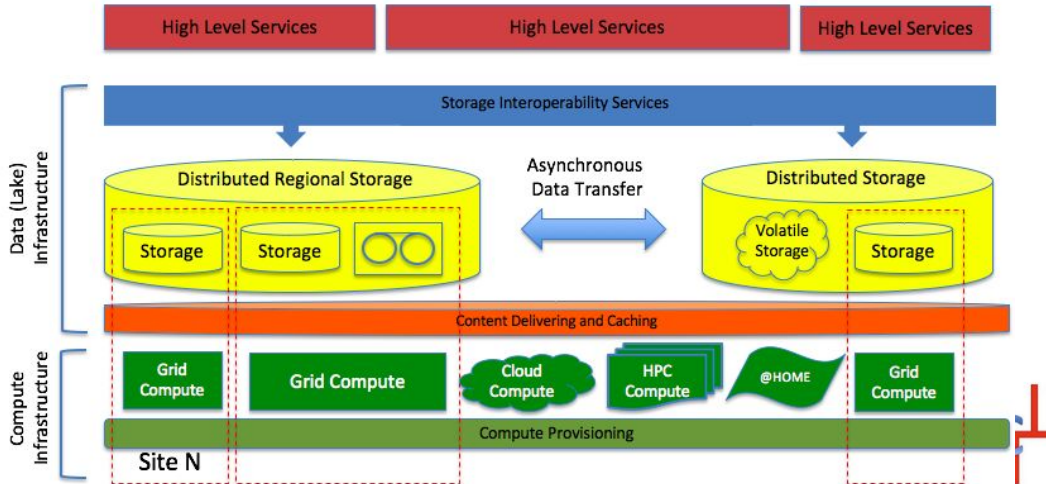
- 1) Only links involving LHCONE are shown
- 2) LHCOPN links are not shown on this diagram
- 3) For more coloration see: "Integrating the LHCONE Map" at <https://www.hepforge.com/docs/2024/03/2024-03-20-integrating-the-lhcone-map/>
- 4) GEANT and CANARIE have shutdown the peering between their VRF and KfAE, as a result of the Ukraine war.

Data Lake

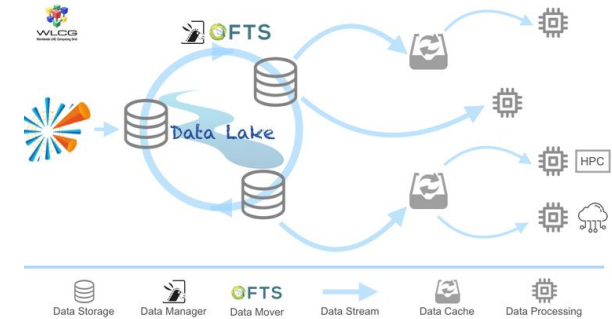
Moving to a network-centric model

CPU and storage not necessarily co-located:

- need fast and reliable WAN
- Storage might offer different QoS classes
- caching for remote compute Infrastructure



Data Lake model



Fewer number of site operating storage services -> **less data replication**

<https://doi.org/10.1051/epjconf/201921404024>

OPENcoasts Coastal forecasts



Coastal forecasts on-demand. Accurate and timely predictions on water conditions.

- Water levels and velocities, wave characteristics
- Forecasts using the SCHISM model

Atmosphere forcings:

- GFS (NOAA)
- ARPEGE (MétéoFrance)

Ocean forcings:

- PRISM2017 (LNEC)
- FES2014 (LEGOS)
- CMEMS (Copernicus)

Wave forcings:

- North Atlantic WW3

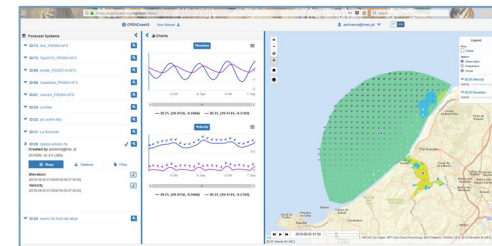
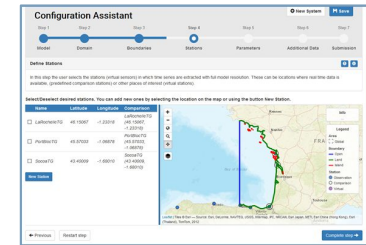
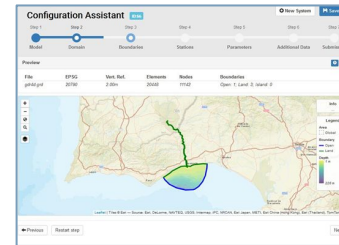
Data:

- Buoy sensors
- Satellite images
- Weather info

onweek.org/agenda/

Summary

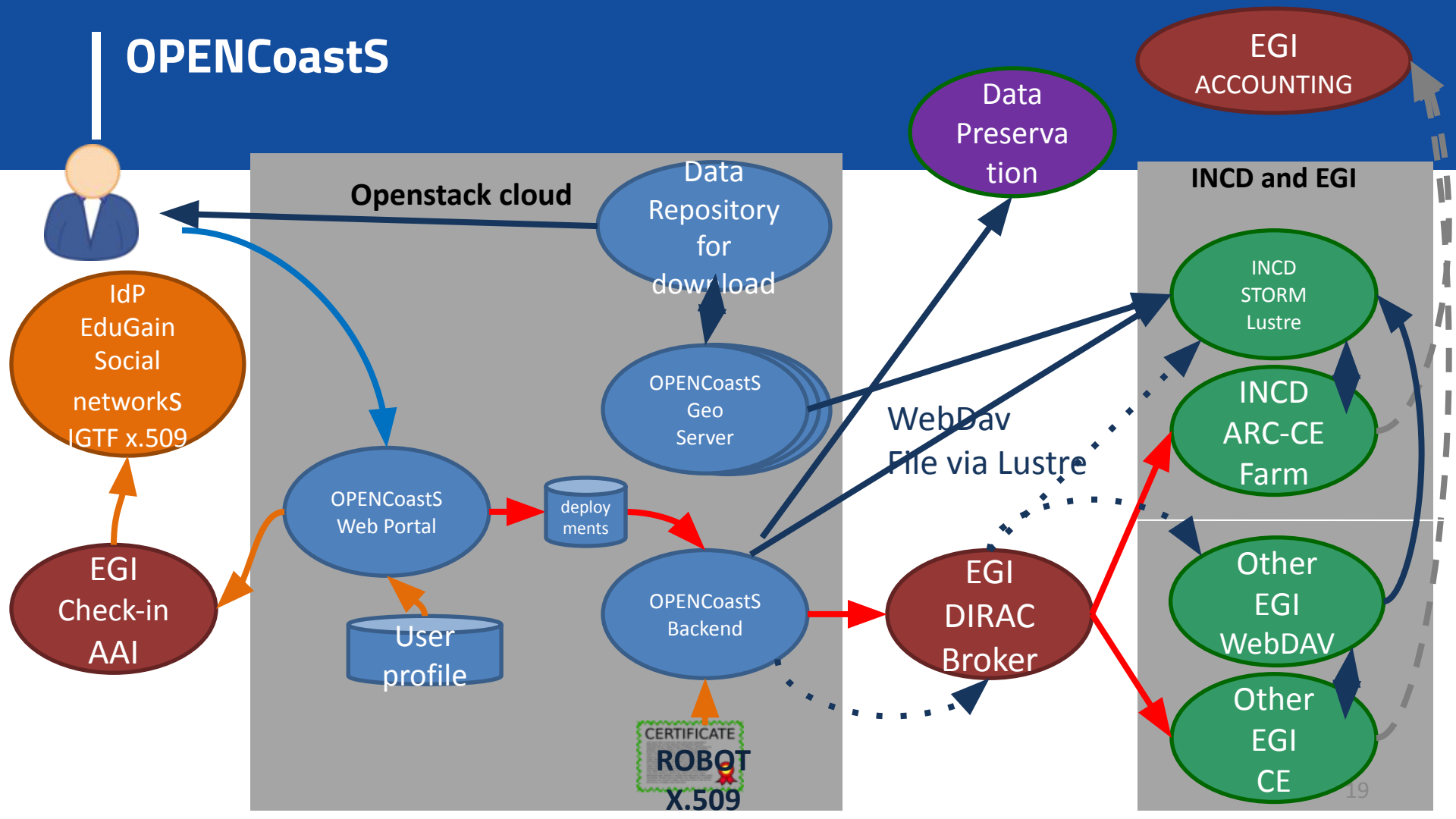
The OPENCoasts service assembles on-demand circulation forecast systems for selected coastal areas and keeps them running operationally for a period defined by the user. This service generates daily forecasts of water levels and vertically averaged velocities over the region of interest for 48 hours, based on numerical simulations of the relevant physical processes.



<https://opencoasts.ncg.ingrid.pt>



OPENCoastS



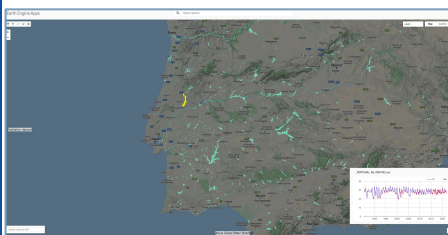
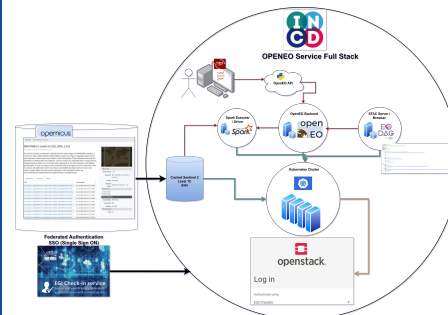
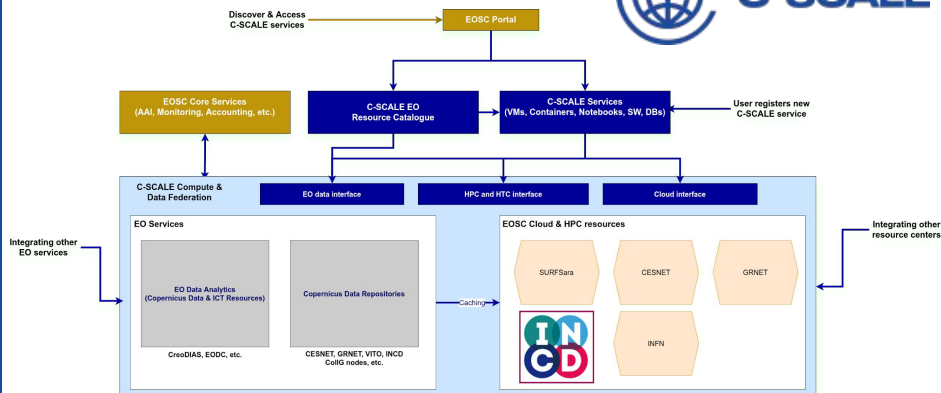
C-Scale Copernicus



European Open Science Cloud project.

C-Scale federate European EO infrastructure services:

- Copernicus DIAS and others.
- Capitalise on the European Open Science Cloud (EOSC) capacity and capabilities.
- Support Copernicus research and operations with large and easily accessible European computing environments.



Aqua Monitor detects how the Earth's surface water has changed during the last 30 years.

Changes are detected in real-time using satellite imagery for any place on Earth.

Porting the application from the Google Earth Engine platform to the open C-SCALE infrastructure, providing an interactive (zoomable) map that displays land use changes (wet vs dry).

Relies on the top-of-atmosphere reflectance images from Landsat 4,5, 7, and 8 and will be extended to use Sentinel-2 MSI Level-1C data.

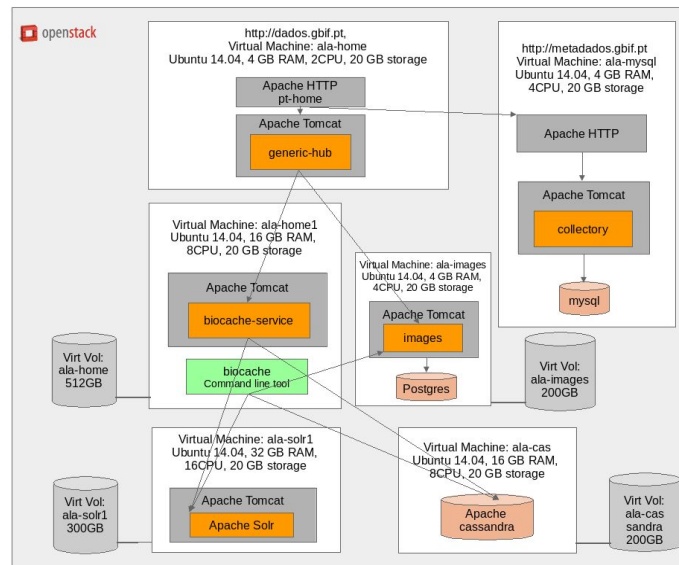
Biodiversity



Integration of Iberian biodiversity information in the Global Biodiversity Information Facility.

- Joining the Spanish and Portuguese GBIF nodes
- Integration of GBIF nodes in EOSC
- Enabling combined analysis of data
- Implement resiliency for both national nodes

Collaboration with the GBIF nodes of Portugal (ISA) and Spain.



Artificial Intelligence: AI4EOSC

- Advanced services for Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) models and applications in the EOSC.
- Event-driven data processing based on serverless computing.
- Agrometeorological forecasts
- Integrated plant protection scenario
- Automated Thermography
- LIP contribution:
 - Software, services and applications quality, data FAIRness
 - Integration of udocker for serverless computing

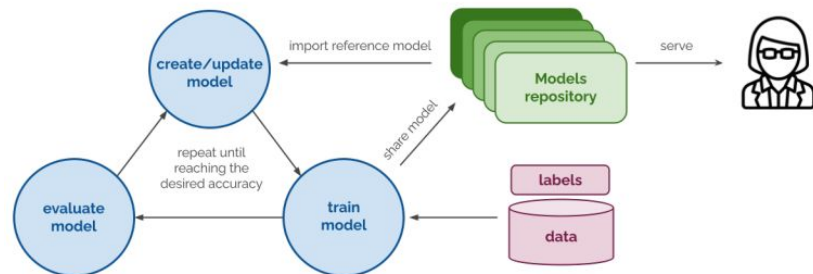
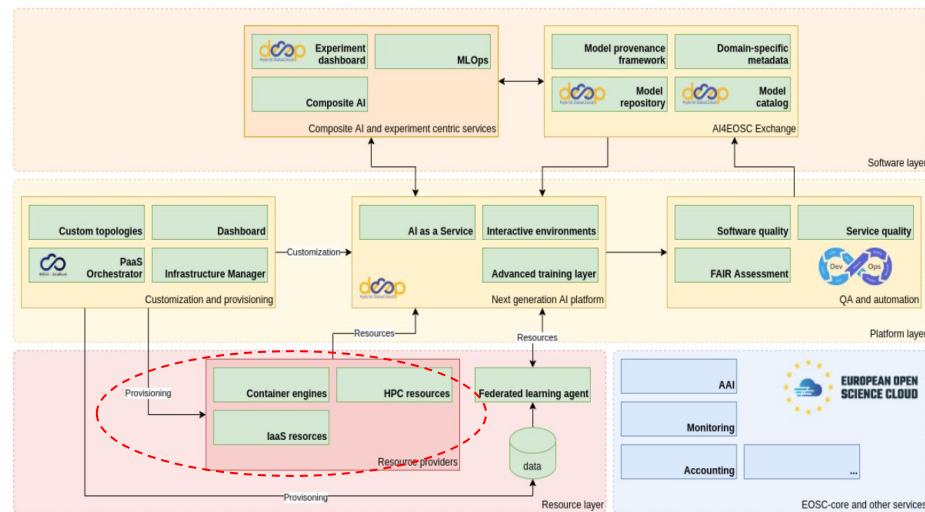
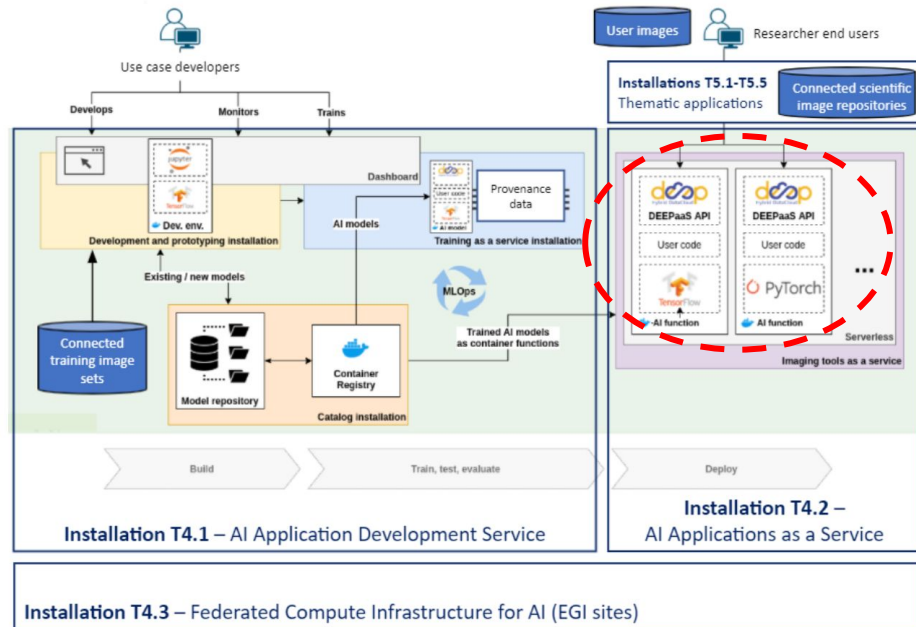


Figure 1: Machine learning application development lifecycle.



Artificial Intelligence: iMagine

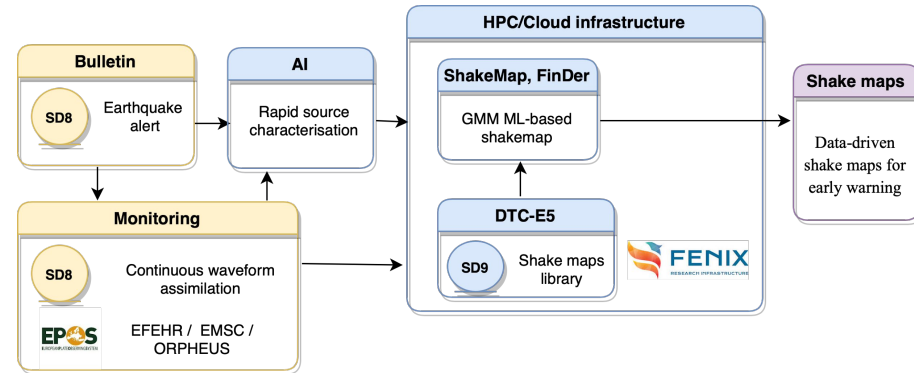
- Imaging data and services for aquatic science
- Ocean warming, and acidification
- Litter and oil spills monitoring of water surfaces
- Marine biomass estimation and preservation through real-time monitoring
- Coastal ecosystems, and beach-related human activities monitorization and analysis
- LIP contribution:
 - DEEP AI Application Development Service
- INCD contribution:
 - Federated Compute Infrastructure for AI and support



Digital Twins: DT-GEO

- Digital Twin for GEophysical extremes (DT-GEO) that aims to analyse and forecast the impact of tsunamis, earthquakes, volcanoes, and anthropogenic seismicity.
- LIP contribution:
 - Workflows and support for deployment of components in HPC and Cloud resources

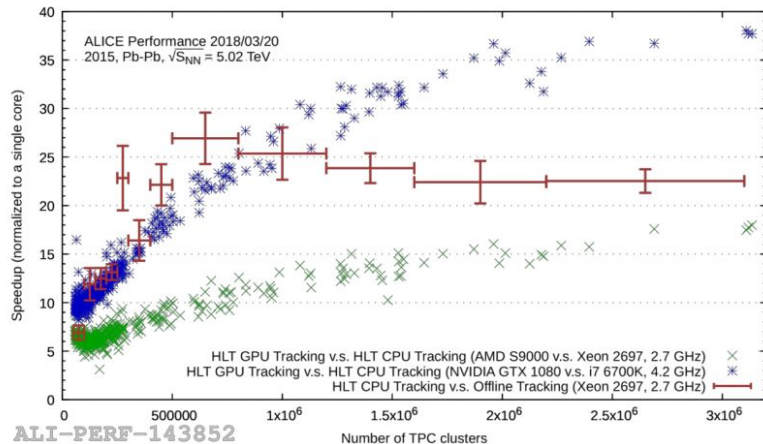
Rapid event and shaking characterisation



Software

The real HL-LHC computing challenge is a **software problem**:

- **Moore's law** is still there in the number of transistors
 - But not in an easily usable form – many cores, specialized co-processors (e.g. vector units), GPUs, etc.
 - (Most of) Today's software is not efficient on these processors (**legacy code x86_64**)



- use new types of resources (HPC, GPU, other non-x86)
 - Implies significant re-engineering & re-writing of core application software
 - Most of the code developed by physicians over the last 10y
 - This is a non-trivial and long-term proposition

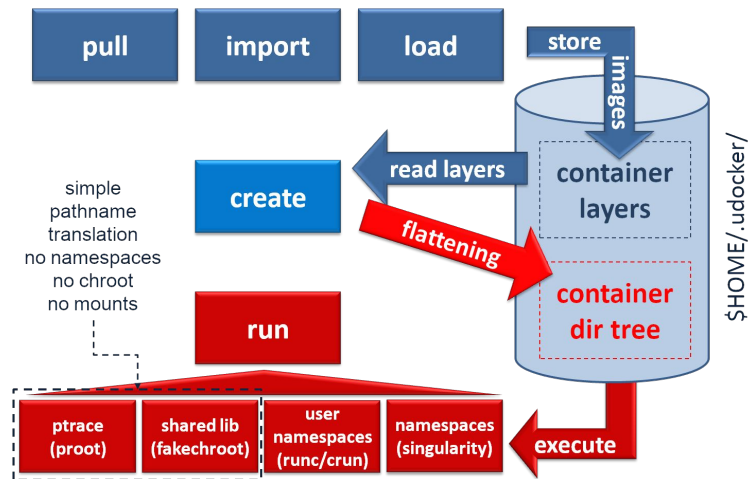
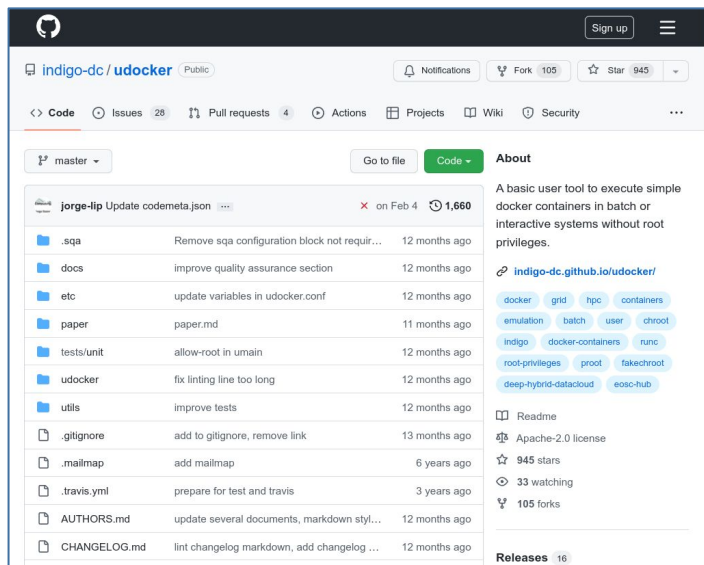
udocker



User tool to execute docker containers in user space. Developed at LIP:

- Fully user space.
- No root privileges required to use or install.
- Does not require compilation.
- Download and execution of docker containers by non-privileged users.
- Suitable for Linux batch systems and interactive clusters managed by other entities such as grid infrastructures.
- Does not require Linux namespaces.

<https://github.com/indigo-dc/udocker>

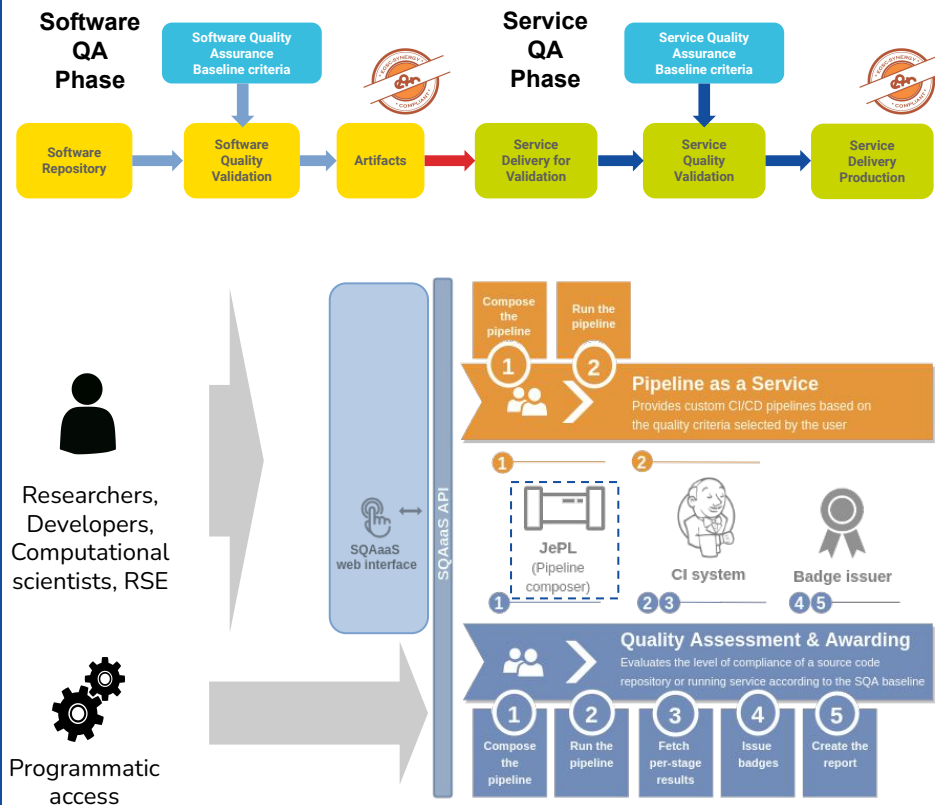


Software from LIP SQAaaS

Quality Assurance as-a-Service platform (SQAaaS)

- Enables the on-demand creation of CI/CD pipelines making quality verification and validation easily accessible to developers.
 - The **Pipeline as a Service** building block allows you to compose and test customized CI/CD pipelines in accordance with reference criteria.
 - The **Quality Assessment & Awarding** building block analyses, the level of compliance to the quality baselines.
- Integrates a wide range of quality verification tools that are made easily available through a friendly web interface.

<https://sqaas.eosc-synergy.eu>



SQA baseline dynamic stages	Environment Setup	qc_style o3api	qc_coverage o3api	qc_functional o3api	qc_security o3api	qc_doc o3api	Push Images to Docker Registry	Docker Compose cleanup
14s	5s	1min 43s	23s	1min 50s	10s	1min 14s	7s	5s
14s	5s	1min 43s	23s	1min 50s	10s	1min 14s	7s	5s

Final remarks

- HEP is highly compute and data intensive and has very large specific requirements
 - Computing infrastructures and support are both essential and an integral part of HEP activities and they have been extremely beneficial to the wider research community and society.
- The infrastructure capacity for HEP activities requires huge amounts of capacity and human resources
 - Lack of human resources it's playing a key factor on today's deployment even with full service automatization. Use of AI may be a solution for some cases.
- dedicated facilities such as the WLCG are needed across multiple domains and technologies
 - They are very specialized and suitable to other similar uses cases and they provide useful services that can be used by other communities
- Better code development and planning are required in scientific computing
 - Use of methodologies based on good practices reduce programing effort cost



LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Thanks!

**Discovery
through
science**

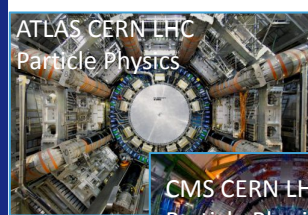
**Innovation
through
technology**

**Sharing
with People**

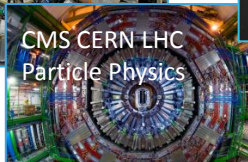


Biomolecular SIMulations Research Group (BioSIM) / UCIBIO
 Business Research Unit (BRU) / ISCTE
 Concepção e Desenvolvimento de Sistemas de Informação (CDSI) / ISCTE
 Centro de Estudos de Doenças Crónicas (CEDOC) / UNL
 Centro de Engenharia e Tecnologia Naval e Oceânica (CENTEC) / U.Lisboa
 Centro de Ecologia Evolução e Alterações Ambientais (CE3C) / U.Lisboa
 Centro de Ecologia Funcional (CEF) / U.C
 Centro de Física do Porto (CFP)
 Centro de Investigação em Biomedicina (CBMR) / UALG
 Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR) / U.Porto
 Fundação Champalimaud
 Instituto de Biosistemas e Ciências Integrativas (BioISI) / U.Lisboa
 Centro de Neurociências e Biologia Celular (CNBC) / U.Coimbra
 Departamento de Informática / FCUL / U.Lisboa
 Departamento de Informática / FCT-UNL
 Departamento de Engenharia Civil / U.Aveiro
 Departamento de Física / IST / U.Lisboa
 Departamento de Física / FCUL / U.Lisboa
 Departamento de Física / U.Minho
 Departamento de Engenharia Informática / U.Coimbra
 Departamento de Informática / U.Porto
 Departamento de Química / U.Aveiro
 Faculdade de Ciências Sociais e Humanas / UNL
 Faculdade de Letras / U.Porto
 Faculdade de Medicina / U.Lisboa
 Fundação para a Ciência e a Tecnologia - Unidade FCCN (FCT-FCCN)
 Nova School of Science and Technology (FCT/UNL)
 Institute for Bioengineering and Biosciences (IBB) / U.Lisboa
 Instituto de Biologia Experimental e Tecnológica (IBET/ITQB)
 Instituto de Geografia e Ordenamento do Território (IGOT) / U.Lisboa
 Instituto de Ciência e Inovação em Engenharia (INEGI)
 Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento (INESC-ID)
 Instituto Nacional de Saúde Doutor Ricardo Jorge (INSA)
 Instituto de Ciências Sociais da Universidade de Lisboa (ICS) / U.Lisboa
 Instituto de Saúde Ambiental (ISAMB) / U.Lisboa
 Instituto Tecnológico e Nuclear (ITN) / U.Lisboa
 Instituto de Plasmas e Fusão Nuclear (IPFN) / U.Lisboa
 Instituto Superior de Agronomia (ISA)
 Instituto Superior de Engenharia de Coimbra (ISEC)
 Instituto Superior de Engenharia de Lisboa (ISEL)
 Instituto Universitário de Ciências Psicológicas, Sociais e da Vida (ISPA)
 Júnior Empresas do Instituto Superior Técnico (JUNITEC)
 Computer Science and Engineering Research Centre (LASIGE) / U.Lisboa
 Laboratório de Instrumentação e Física Experimental de Partículas (LIP)
 Laboratório Nacional de Engenharia Civil (LNEC)
 Nova School of Business and Economics (Nova SBE) / UNL
 Rede de Química e Tecnologia (REQUIMTE) / UCIBIO
 Sociedade Portuguesa de Botânica (SPBOTANICA)
 Universidade Aberta (U.Aberta)
 Applied Molecular Biosciences Unit / UCIBIO
 Portuguese node of the Global Biodiversity Network (GBIF)
 Portuguese Coastal Monitoring Network (CoastNET)
 Portuguese Biological Data Network (BIODATA.PT)

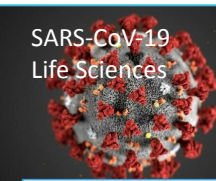
Users from 53
 research units
 and
 organisations
 in 2022



ATLAS CERN LHC
Particle Physics



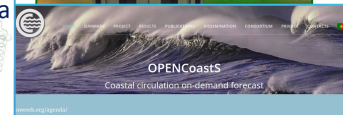
CMS CERN LHC
Particle Physics



SARS-CoV-19
Life Sciences



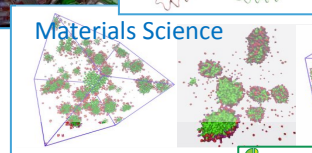
AUGER
Astroparticle Physics
Cosmic Rays



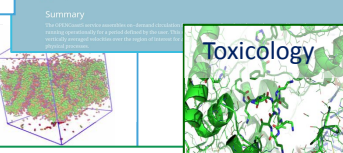
OPENCOSTS
Coastal circulation on-demand forecast



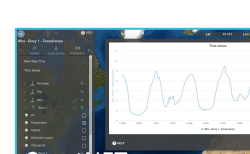
Chromatin & Epigenetics,
Clinical Pharmacology,
Cancer Signaling,
Computational Biology



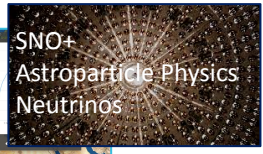
Materials Science



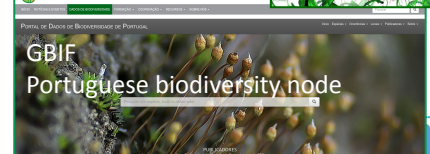
Toxicology



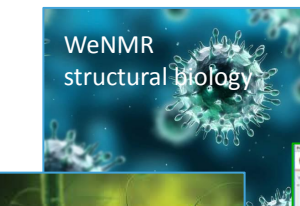
CoastNET
Coastal monitoring



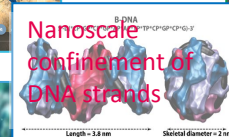
SNO+
Astroparticle Physics
Neutrinos



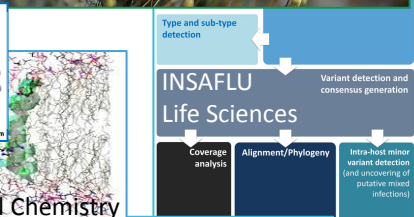
GBIF
Portuguese biodiversity node



WeNMR
structural biology



Nanoscale
confinement of
DNA strands



INSAFLU
Life Sciences



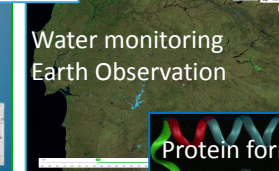
SCIPION
Computational Chemistry



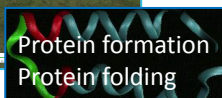
BIOMED VO
Life
Sciences/Genetics



INSTRUCT
Life Sciences



Water monitoring
Earth Observation



Protein formation
Protein folding

EuroCC HPC Competence Center

EuroHPC support project (2020-2022-2025)

High Performance Computing National Competence Center in EuroCC.

- Boost European HPC knowledge
- European network of 33 national HPC competence centres (NCCs)
- Bridge the skills gaps and promote cooperation

Areas

- Technology transfer
- Training and skills development
- **Awareness and collaboration**
- Collaboration with industry
- Access to expertise and knowledge

EUROCC PORTUGAL

ABOUT EUROCC HPC IN PORTUGAL NEWS AND EVENTS TRAINING AND SUPPORT CONTACTS

EURO

Competence Centers for HPC

A source of information, training and collaboration.

News and Events

Registration to the 3rd edition of LIP Data Science is now open

events | news

From June 27 to July 1, Coimbra hosts the 3rd edition of the Data Science School and Symposium, organized by LIP. All PhD students and young post-docs in physics and engineering are invited to participate!

ISC 2022: Europe's biggest HPC event is about to start

events | news

Hamburg will host ISC High Performance 2022 from May 29 to June 2. Trends and major developments in HPC, machine learning and data analytics are the central topics of this years edition. More than 60 countries and three thousand attendees will be involved.

IBERGRID 2022: Delivering Innovative Computing and Data Services for Research

events | news

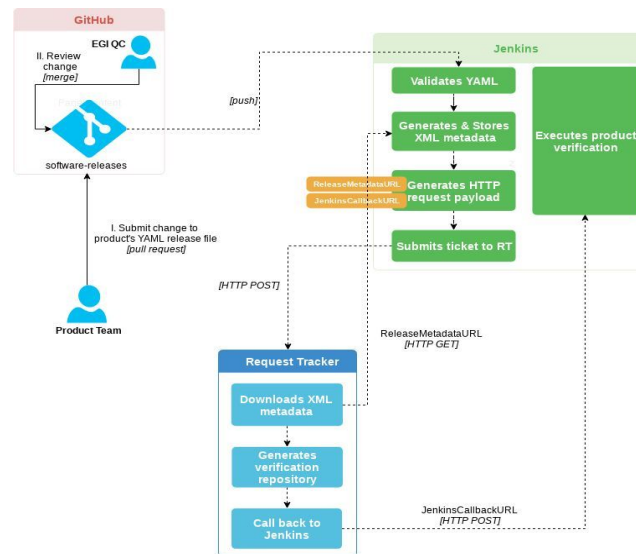
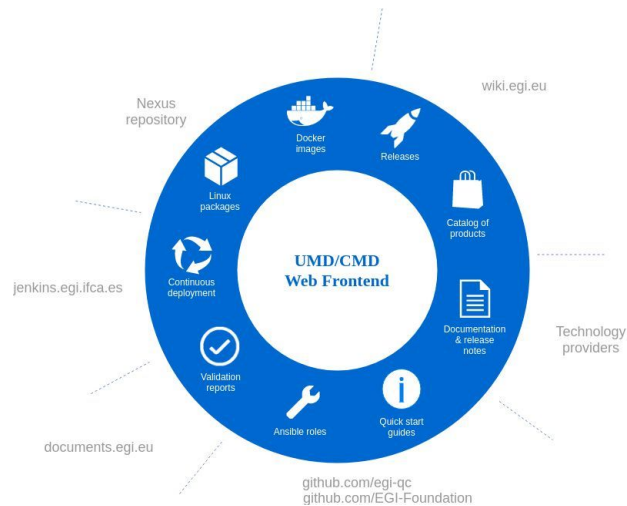
We are happy to announce that the 11th edition of the Iberian Grid Conference will take place in Faro, at the University of Algarve, from 10th to 14th October 2022.

<https://eurocc.fccn.pt> / contacto@eurocc.fccn.pt

Software quality

Quality assurance for the EGI middleware distributions for grid (UMD) and cloud (CMD).

- Production
 - Manage the EGI software validation process.
 - Software validation of products to be released as part of CMD and UMD distributions.
 - Automated validation in isolation environments and piloting at selected sites.
- Innovation
 - New streamlined validation process.
 - New repositories with added capabilities.
 - New frontend.



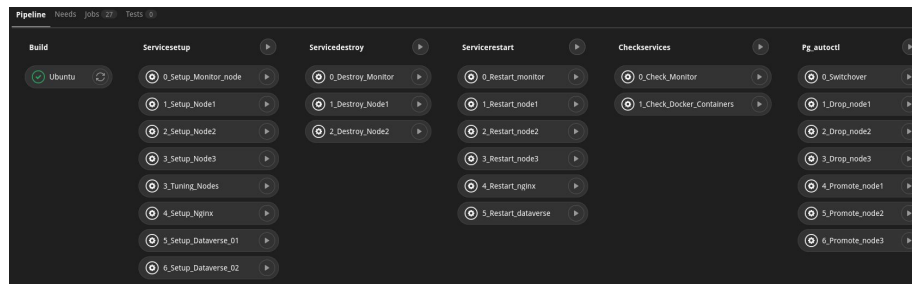
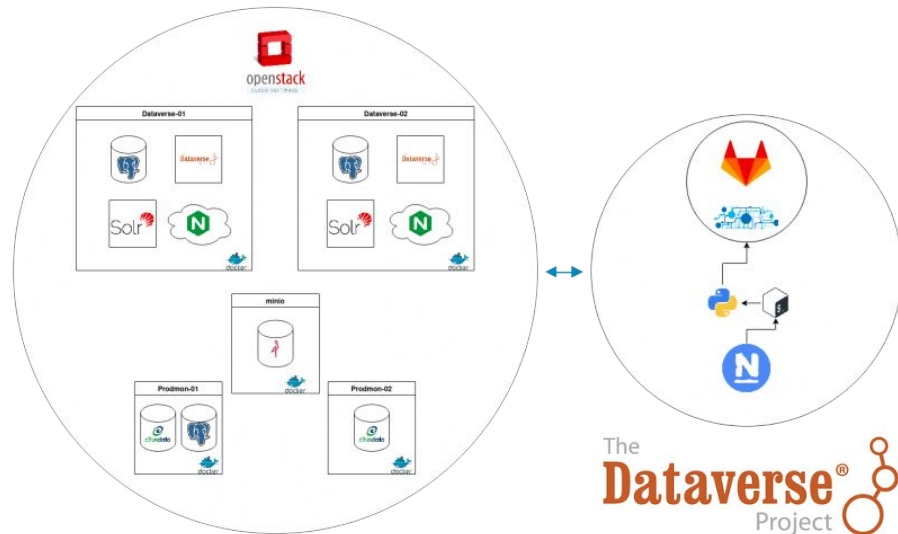
Research data Repositories



FCT contract, performed under INCD activities.

Feasibility study for a national catchall data repository aligned with open science and FAIR data principles.

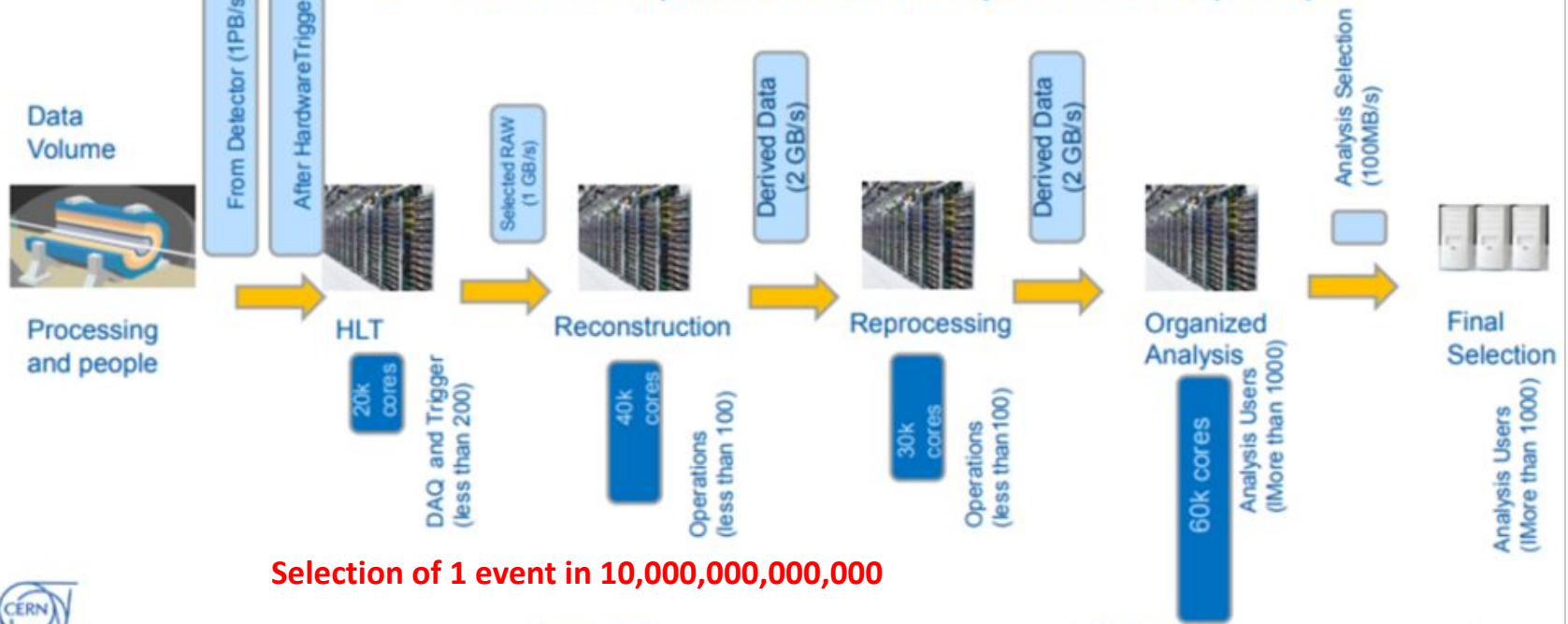
- Leverage EOSC-Synergy work on thematic data repositories and FAIR quality indicators.
- Productization and automation of Dataverse based data repositories.
- Address resiliency, redundancy and data recovery aspects.
- Integration with federated identity e.g. Ciência ID, RCTSaaI.
- Integration with permanent identifiers.



Data Analysis at the LHC

The process to transform raw data into useful physics datasets

- This is a complicated series of steps at the LHC (Run2)



Deep Learning / Machine Learning as a Service

Use of specialized hardware:

- GPUs, Low latency interconnects



On demand

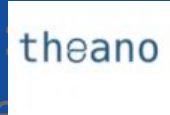
Virtualized environments: (u)Docker, LXD, QEMU/KVM

Docker Orchestration: Mesos/Marathon, Kubernetes



Stack theano, TensorFlow, no ...

Apache Spark/Flink



DevOps for containers

