jornadas FCCN

INCD

Utilização dos recursos da Google cloud na INCD

jornadas.fccn.pt

fct Fundação para a Ciência e a Tecnologia

FCCN

arditi agência regional para o desenvolvimento da investigação tecnologia e inovação

SIH

UNIVERSIDADE da MADEIRA

National Distributed Computing Infrastructure

Services: scientific computing, data processing and other data oriented services

Target: scientific and academic community, infrastructures, R&I projects, SMEs

Promote: shared resources, advanced computing and data services for research

Interface: international digital infrastructures and initiatives (EGI, IBERGRID, WLCG, EOSC)

Cloud Computing

cloud computing

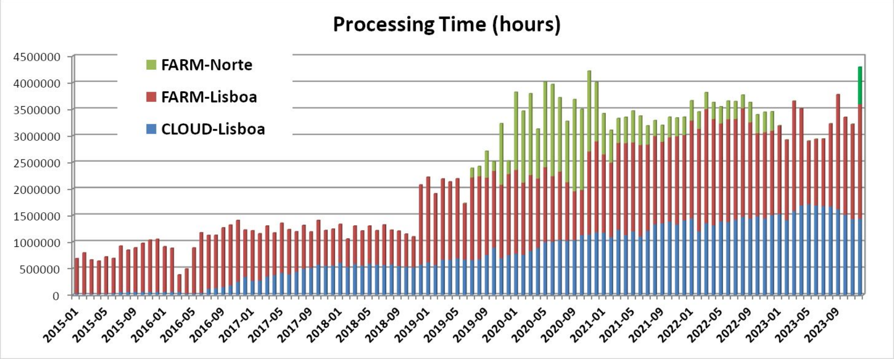
HTC Computing

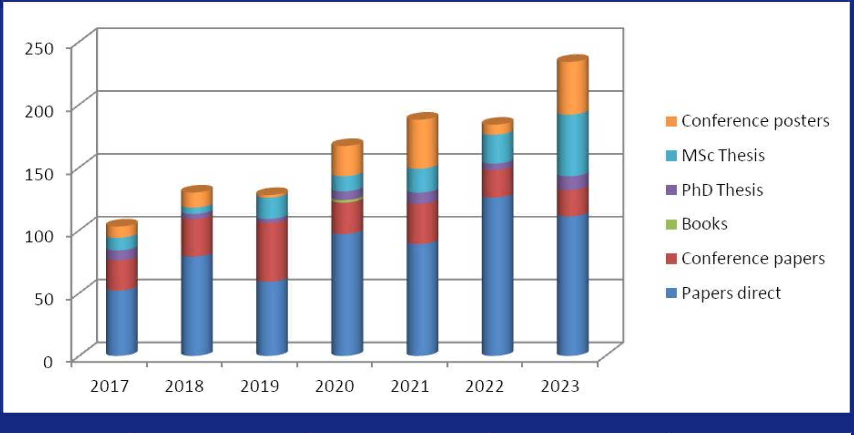
high throughput computing (GRID)

HPC Computing

high performance computing

Global INCD usage and metrics





	2017	2018	2019	2020	2021	2022	2023	Total
Publications	52	79	59	97	89	126	111	613
Proceedings	24	30	47	25	32	22	21	201
Books	0	0	0	2	0	0	0	2
PHD thesis	8	4	3	7	9	5	11	47
MSc thesis	10	5	17	12	19	23	49	135
Posters	9	12	2	24	39	8	42	136
Patents					2	1	0	3
Datasets and open source software				1	2	19	11	33
Total	**103**	**130**	**128**	**168**	**192**	**204**	**245**	**1170**

The EOSC-Future call for commercial cloud redistribution

Digital service aggregators (e.g. non-profit entities, NRENs, RIs and e-Infrastructures, HPC centres, etc.)

+

OCRE cloud service providers in the same region collaborating on dynamic and creative proposals.

1. Proposals had to demonstrate **a concrete approach to distributing state-of-the-art digital services** (e.g. compute, storage, machine learning, analytics, AI) **via the European Open Science Cloud**.
2. Aiming at mechanisms to potentially drive strategy and relevance for many of the research infrastructures moving forward.

The award:

1. **up to €400 000** (500.000 with VAT) in pre-procured IaaS/PaaS/SaaS from the OCRE cloud provider
2. The call was part of a €4.8M adoption funding programme supported by the EU, through the EOSC Future project (the 1st call distributed €2M)

INCD has been in several EOSC projects:

EOSC-hub · EOSC SYNERGY · EGI-ACE · iMagine · C-SCALE

Was already in OCRE.

Could act as service aggregator !

Potential benefits

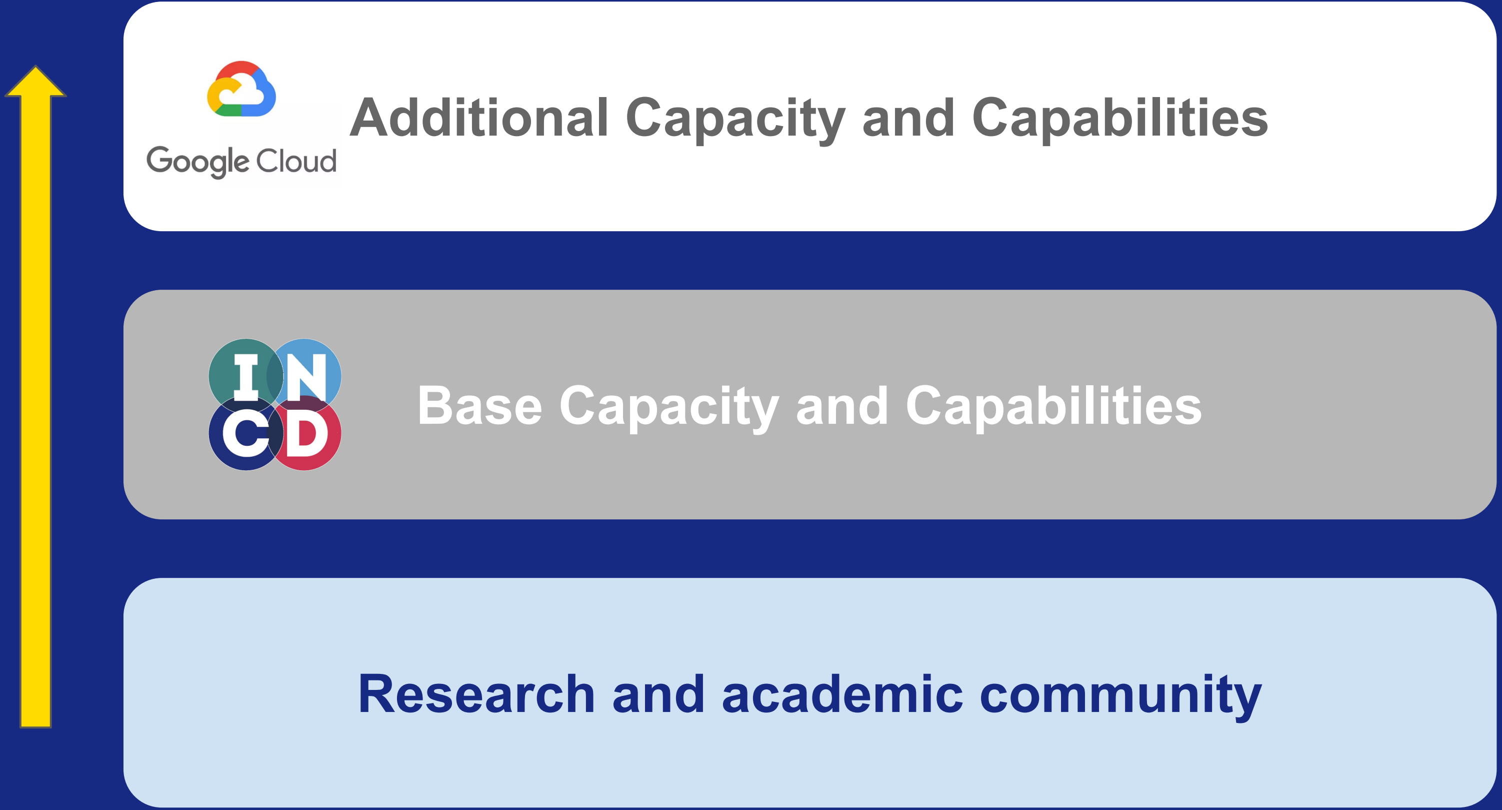
- Access to worldwide distributed resources from Google datacenters
  - Resiliency, geographic coverage
- Extend infrastructure when additional computing resources are required
  - Additional capacity, address a wider range of technical requirements
- Test brand new hardware (TPUs, GPUs, CPUs) available in Google infrastructure
  - Access to expensive and less frequently required therefore hard to justify resources
- Access to GCP added value services
  - Profit from services ready to use that are hard to setup and provide by a small provider
- Facilitate combined use of the INCD services
  - Provide an easier solution for users that in the future might be willing to pay for additional capacity and services

The EOSC-Future call for commercial cloud distribution

CONCEPT: GET MATCHED UP WITH A COMMERCIAL CLOUD PROVIDER!

- Applying required to get matched with a commercial cloud service provider in the region.
- Specifically, from the list of OCRE framework contract holders.

Public-cloud / Private-cloud partnership …



Google Cloud

Additional Capacity and Capabilities

Peaks of usage
Specific services
Specific hardware
Resilience

INCD

Base Capacity and Capabilities

Safe harbour
Sustained usage
Low cost
User support

Research and academic community

Use INCD cloud
Burst into google
Avoid lock-in
Minimise costs

Technical approach

- **Bet on serverless approach**
  - instead of using virtual machines focus on container based services
  - exploit event driven solutions like function as a service, task execution, etc
  - minimise costs

- **Reduce user lock-in and dependencies**
  - users exploit commercial cloud services via INCD
  - use open source solutions
  - INCD as cloud orchestrator

Activities

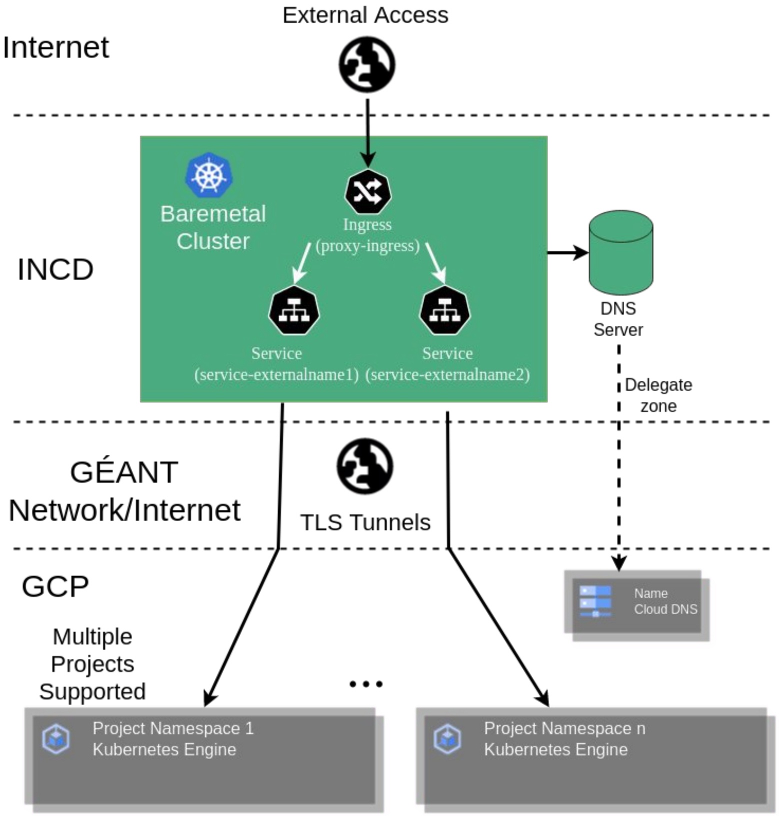
- **Training phase for INCD staff**
  - Provided by Google
- **Understanding several options of use**
  - Goal exploit lower cost solutions
- **Integration**
  - Make access and usage more transparent for the end users
  - Enable users to access Google cloud from the INCD services
  - Integration of the INCD infrastructure with Google cloud
- **Exploitation and Demonstration**

Use case: Applications in the Google Kubernetes Engine

INCD integration with GCP

- **Control traffic through the INCD infrastructure**
  - Endpoints at INCD
- **Gateway under INCD control**
  - Services can move between providers with almost zero downtime
  - Proxies to redirect traffic
- **Published DNS records always under control of the INCD managers**
  - Only DNS records published by INCD are visible
- **Data movement mostly from INCD site to external provider**
  - Minimise future data storage and transfer costs
- **Data policy: allow data to be kept at INCD**
  - Important for customer protection

INCD integration with GCP



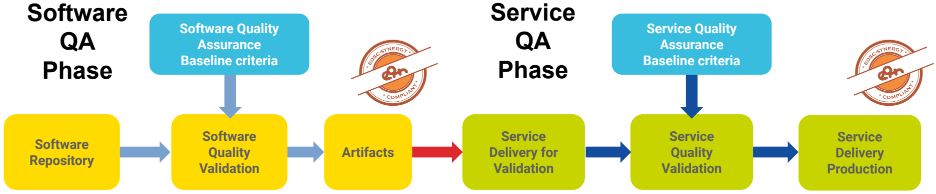
- Endpoints are exposed by INCD
- Endpoints are managed by proxies
- DNS via zones that can be delegated
- Projects can be moved to GKE
- Depending on project access can be:
  - Directed to INCD Kubernetes
  - Directed to Google GKE
- Tunnels between INCD and Google
- Multiple different projects in GKE

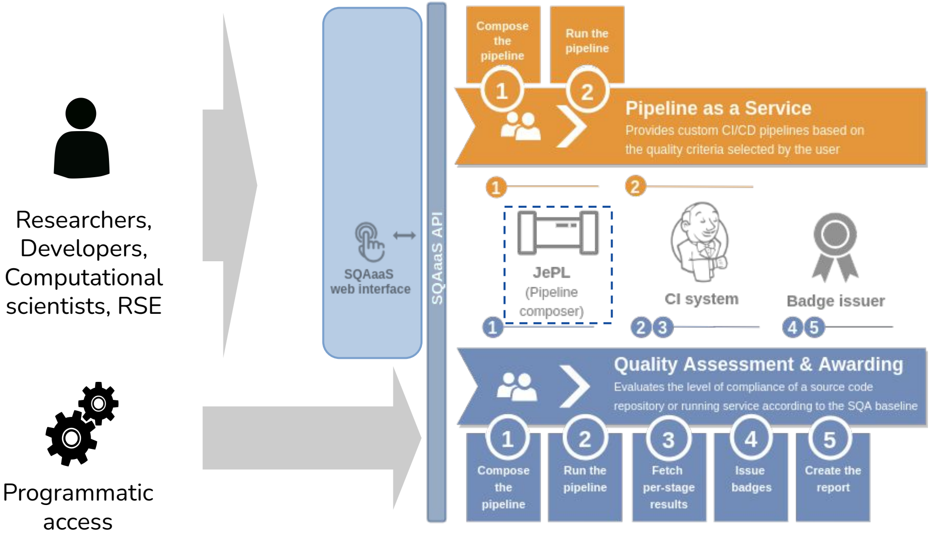
EOSC-Synergy **SQAaaS**

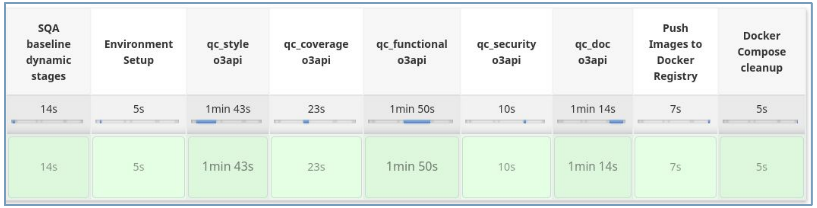
Quality Assurance as-a-Service platform (SQAaaS)

- 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.

- Challenge ⇒ scalability and availability
- **Now hosted at INCD using workers at Google**


Software QA Phase
Software Quality Assurance Baseline criteria
Service QA Phase
Service Quality Assurance Baseline criteria
Software Repository
Software Quality Validation
Artifacts
Service Delivery for Validation
Service Quality Validation
Service Delivery Production


Researchers, Developers, Computational scientists, RSE
Programmatic access
SQAaaS web interface
SQAaaS API
Compose the pipeline
Run the pipeline
1
2
Pipeline as a Service
Provides custom CI/CD pipelines based on the quality criteria selected by the user
JePL (Pipeline composer)
CI system
Badge issuer
Quality Assessment & Awarding
Evaluates the level of compliance of a source code repository or running service according to the SQA baseline
1 Compose the pipeline
2 Run the pipeline
3 Fetch per-stage results
4 Issue badges
5 Create the report



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

EOSC-Synergy
SQA as a Service

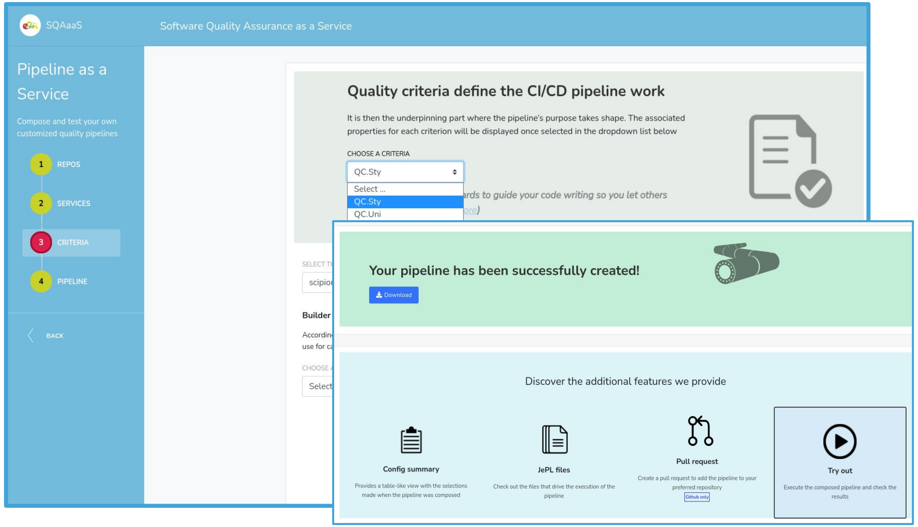
The SQAaaS is provided as a cloud service, making adoption and usage easier.

- No need to deploy and setup the components, Jenkins, API, web, containers.
- No need to create the yaml configurations.
- No need to provide IT resources.
- No need to manage the platform.

Platform for QA of research software & services

- Can issue digital badges to reward and highlight the quality achievements.
- Based on OpenBadges specification.
- Produce detailed quality reports.

https://www.eosc-synergy.eu/technical-areas/quality/



SQAaaS Software Quality Assurance as a Service

Pipeline as a Service

Compose and test your own customized quality pipelines

1 REPOS
2 SERVICES
3 CRITERIA
4 PIPELINE

BACK

Quality criteria define the CI/CD pipeline work

It is then the underpinning part where the pipeline's purpose takes shape. The associated properties for each criterion will be displayed once selected in the dropdown list below

CHOOSE A CRITERIA

QC.Sty

Select ...
QC.Sty
QC.Uni

ards to guide your code writing so you let others
ore)

SELECT TI
scipior

Builder

Accordin
use for ca

CHOOSE A
Select

Your pipeline has been successfully created!

Download

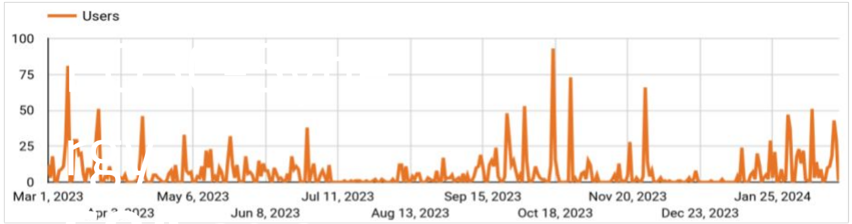
Discover the additional features we provide

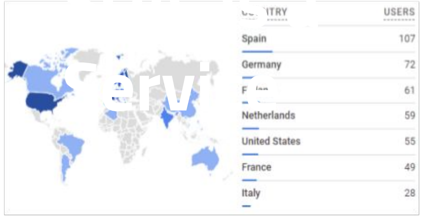
Config summary
Provides a table-like view with the selections made when the pipeline was composed

JePL files
Check out the files that drive the execution of the pipeline

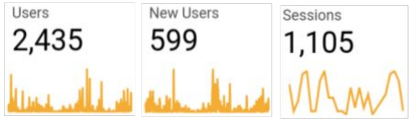
Pull request
Create a pull request to add the pipeline to your preferred repository
Github only

Try out
Execute the composed pipeline and check the results





COUNTRY	USERS
Spain	107
Germany	72
Finland	61
Netherlands	59
United States	55
France	49
Italy	28



Users	New Users	Sessions
2,435	599	1,105

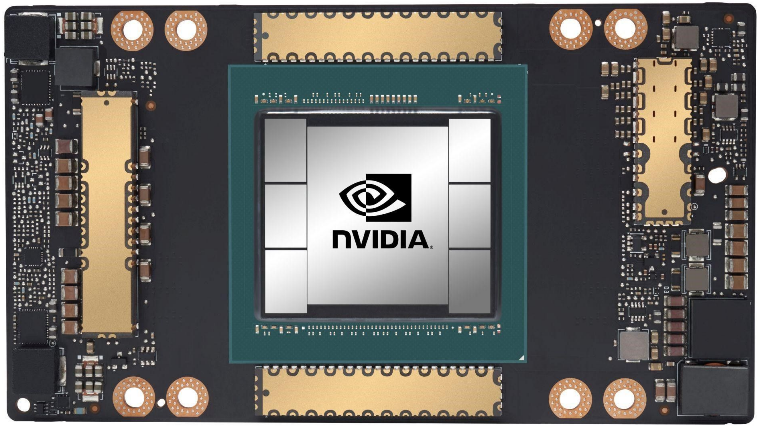
Total badges awarded: 2261
Repositories assessed: 533

Service-Bronze	Software-Bronze	Software-Silver	Software-Gold
898	821	406	136

Use case: GPU usage in the Google Cloud Engine

GPUs

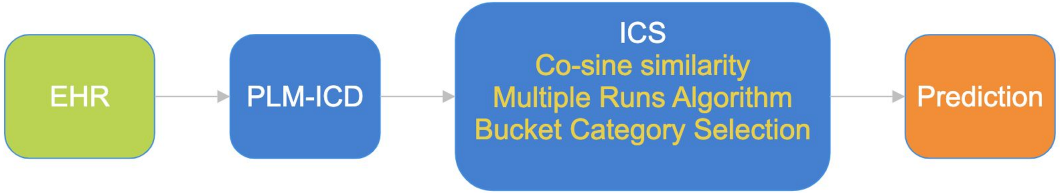
- GPUs are highly demanded
  - Insufficient capacity for requests
  - Accelerated computing particularly in life sciences (gromacs, amber)
  - Artificial Intelligence and ML
- Nvidia A100, V100 and T4
  - Preferably aldo less flexible via the HPC/HTC farm
  - Native access or via containers (apptainer/udocker)
- Still using very old GPUs for education
  - K20 and K40



Codification of clinical episodes in natural language

- The International Classification of Diseases, 10th Revision (ICD-10) has been widely used to classify patient diagnostic information.
  - ○ Coding clinical episodes into ICD-10 codes is a laborious task, usually done by dedicated physicians with specific training.
  - ○ Automatically coding electronic health records (EHR) into diagnosis codes has been challenging for the natural language processing (NLP) community.
  - ○ Project: Instituto Superior de Engenharia de Coimbra + Hospital de Coimbra (Departamento de Doenças Infeciosas)

- Improved cosine method (ICS), combined with a pre-trained language model (PLM-ICD), to increase the number of useful ICD-10 code suggestions, based on the Medical Information Mart for Intensive Care (MIMIC) dataset -IV.
  - ○ Use of PLM-ICD, a deep-learning model for automatic coding of pre-(MIMIC)-IV clinical texts
  - ○ Training for the Dataset (MIMIC)-IV
  - ○ Implementation of a model (PLM-ICD-C) based on PLM BioLM RoBERTa-base-PM-M3-Voc-distill-allign-hf
  - ○ “multiple runs algorithm” to eliminate frequent words that overlap other important words.



EHR → PLM-ICD → ICS → Prediction

ICS
Co-sine similarity
Multiple Runs Algorithm
Bucket Category Selection

Codification of clinical episodes in natural language

- Results:
  - Higher accuracy model that provides better ICD-10 code suggestions.
  - Publication:
    - Silva, H.; Duque, V.; Macedo, M.; Mendes, M. Aiding ICD-10 Encoding of Clinical Health Records Using Improved Text Cosine Similarity and PLM-ICD. Algorithms 2024, 17, 144. https://doi.org/10.3390/a17040144
  - Master's thesis at the Higher Institute of Engineering of Coimbra (ISEC)

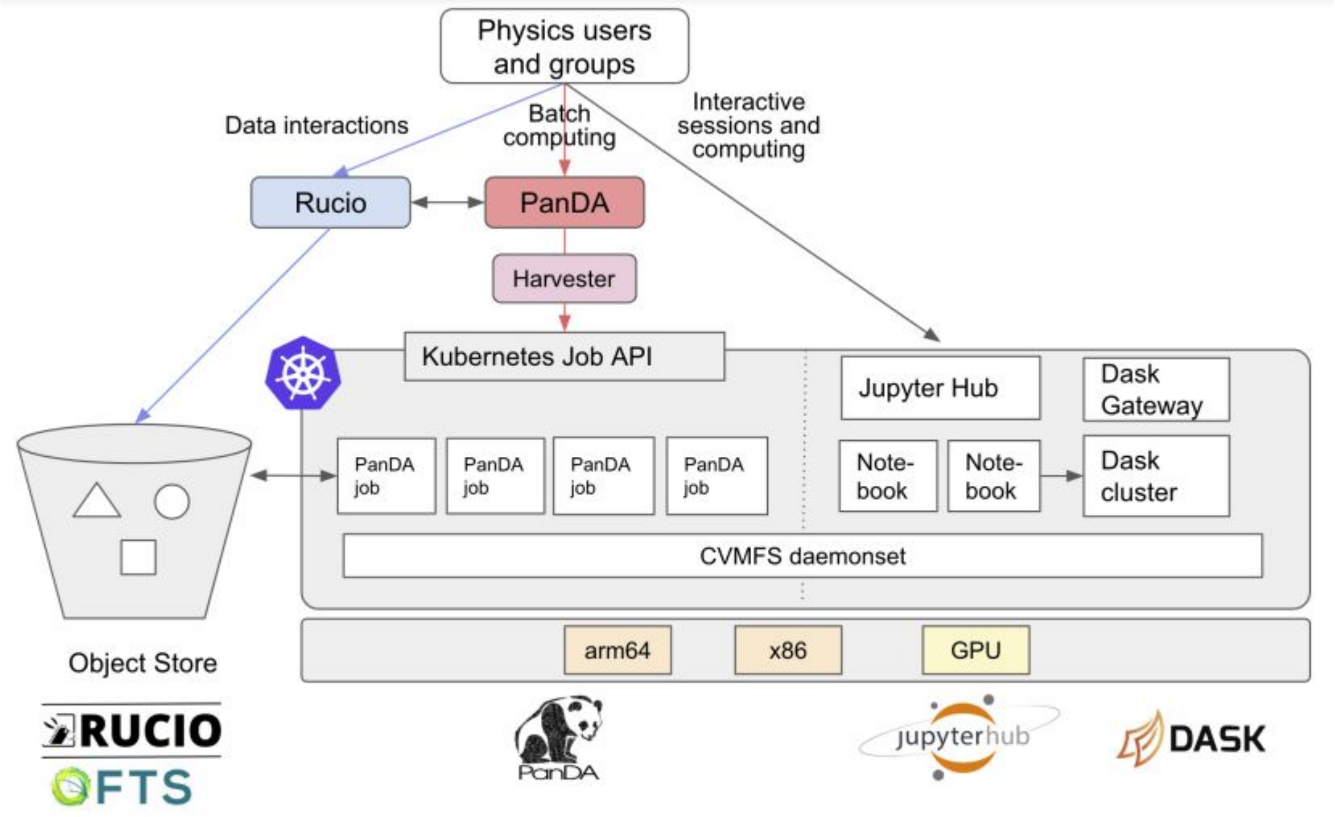
- Computing resources provided by INCD were used
  - HPC systems from INCD
    - GPUs from Cirrus-A Cluster in Lisbon
    - Insufficient capacity available
  - Virtual Machines in the Google Cloud
    - GPUs A100
    - In the context of the INCD + TI Sparkle + Google
    - Equivalent to one month of usage
    - Capacity and system was configured and provisioned by INCD, the end user just used the capacity.
    - Also used for comparison with the INCD owned A100 systems (Google showed better performance)
    - The Google capacity was decisive for the project.

Use case: CERN ATLAS

CERN ATLAS study

- **Evaluation of commercial cloud services by the ATLAS collaboration**
  - *Total Cost of Ownership and Evaluation of Google Cloud Resources for the ATLAS Experiment at the LHC (internal ATLAS note of 27 February 2024)*
  - *Operational Experience and R&D results using the Google Cloud for High Energy Physics in the ATLAS experiment: https://arxiv.org/pdf/2403.15873.pdf*
- **Conclusions:**
  - Network high costs for certain ATLAS workflows (up to 54% of monthly costs)
  - Storage: $20 / TB month Network: $45 to $85 / TB
  - The cost of storage and network can vary considerably depending on main activity
  - Aggressive purging of data may be required to minimise costs
  - ATLAS Google site was very stable
  - Spot instance eviction rates up to 5%
  - Resource bursting was highly successful
  - Scaling to more than 100k concurrently running vCPUs in PanDA was demonstrated
  - Subscription agreement model is essential to contain costs

CERN ATLAS study



Operational Experience and R&D results using the Google Cloud for High Energy Physics in the ATLAS experiment

Final remarks

Initial cost analysis for the INCD use cases

Network pricing

- GCP Regional External Outbound Networking
  - 8.52€ / TB
  - Waiver discount of 100% => 0€ / TB
- GCP NAT data processing
  - 42.60€ / TB
- Cloud Balancer Forwarding Rule
  - 16.64€ / month

Compute pricing

- Compute Engine VMs cost per core
  - 23.77€ / month
- Nvidia Tesla V100
  - 1697.47€ / month
- Compute Engine VMs memory
  - 3.19€ / GB month

GKE pricing

- GKE Anthos (Kubernetes service mesh solution)
  - 10.94€ / month
- GKE Autopilot Pod mCPU 1K Requests
  - 32.62€ / month
- GKE Autopilot Pod Memory Requests
  - 3.61€ / GB month
- GKE Autopilot Pod Ephemeral Storage
  - 41.07€ / GB month

Storage pricing

- Cloud Logging Storage
  - 473.37€ / TB month
  - Free tier until 50GB
- Compute Engine Storage Persistent Disk
  - 41.66€ / TB month
- GKE Balanced Persistent Disk
  - 73.22€ / TB month

Discounts of 10% apply the costs shown

Next steps

- **Increase the use cases / workloads in the google cloud via the INCD brokering**
  - add further use cases and increase workload over the next months
  - exploit other services e.g. execution of tasks and cloud functions (FaaS)
- **Pave a way for a longer term partnership**
  - researchers will be able to use the INCD cloud services for their day-to-day needs as usual
  - and complement with Google Cloud for added capacity and capabilities
- **The business model may offer**
  - a cost effective way to leverage both the INCD and Google Cloud services
  - a partnership that will be advantageous for all interested parties

https://www.incd.pt INCD helpdesk@incd.pt

Thank you

Google Cloud SPARKLE