INFN Approach in Handling Sensitive Data

Managing & Processing Sensitive Data

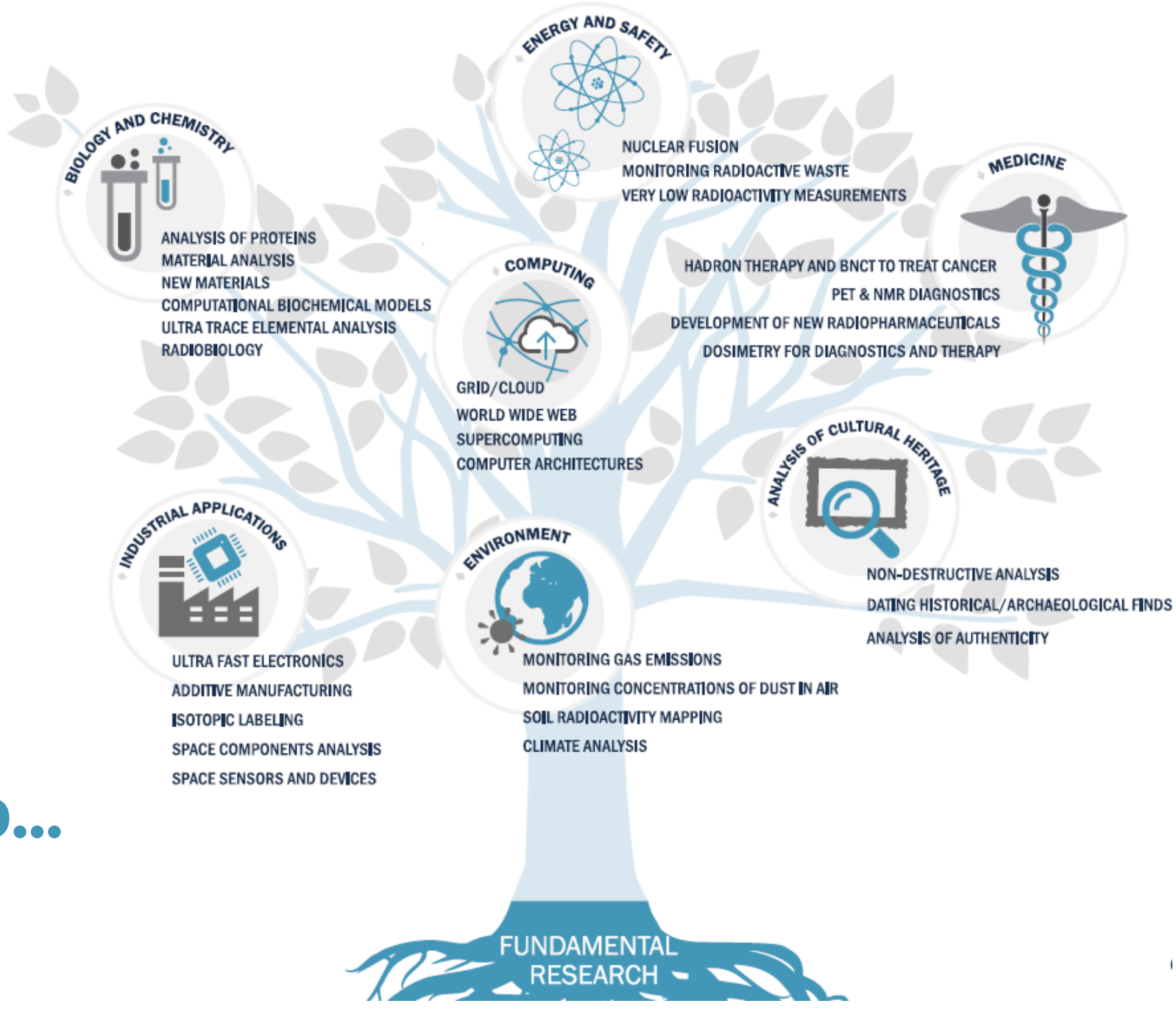
EGI2024 Conference

Alessandro Costantini

Alessandro.costantini@cnaf.infn.it

The context

From physics to...



INFN computing and big data management infrastructure

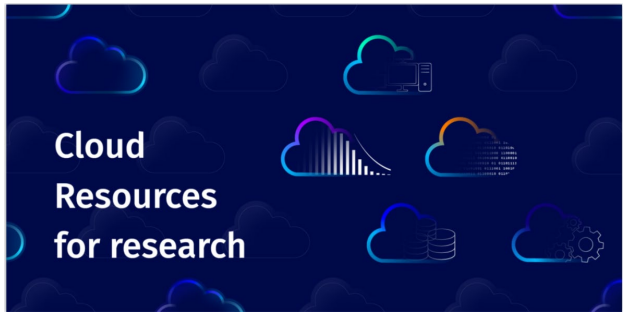
- A long tradition in **state-of-the-art distributed IT technologies and solutions**, from the first small clusters to Grid and Cloud-based computing.
- INFN is not interested in computing per-se, but as an essential way to **support its research and mission**.
- INFN operates Grid and Cloud services based on its own:
  - 1 large national center, at CNAF (Bologna) – **with an area certified ISO/IEC 27001, 27017, 27018**
  - 9 medium size centers
- All the INFN centers are connected through 10-100 Gbit/s dedicated links via the GARR network.
- Collectively, our distributed infrastructure currently offers about **140,000 CPU cores, 120PB of enterprise-level disk space, 100PB of tape storage**

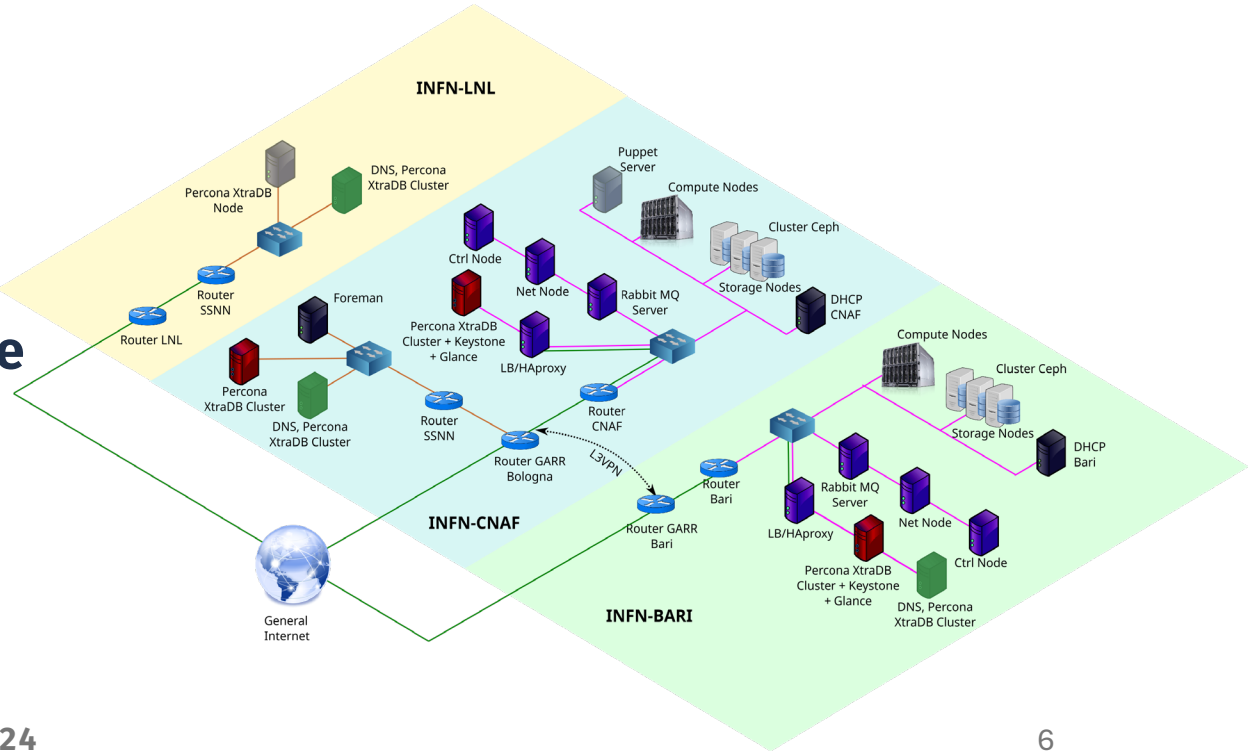
Computing related projects at INFN

- **Scientific Computing within INFN** originally driven by the needs of its own theoretical and experimental communities.
- Being at the forefront of computing in research seeded **many projects with a much broader scope**.
- The **key overall driver** was always to let our users **effectively exploit all available resources and technologies**.
- In 2000-2005 ten main international centers selected to host the **Worldwide LHC Computing Grid (WLCG)**
- Then came the **GRID**, **the Cloud**, …
- All centers still operational. Their **size has increased ~100x** since then, with an **interconnectivity** (thanks to the **GARR-X network**) **up to several 100s Gbps.**

INFN Cloud - https://www.cloud.infn.it

- The **starting point** for a **National Data Lake** for research and beyond, building on (existing | renewed | new) e-Infrastructures.
- The **base of the evolution** of the INFN Distributed Computing vision.
- Built on a **middleware layer** running on top of **federated** clouds, decoupling physical and logical views via a **service composition** mechanism.
- In perspective, it will be the **Italian Node of the** for **HL-LHC**.
- **More than 2y of production service**.





EPIC Cloud

The INFN Cloud region dedicated to sensitive data management

EPIC
Enhanced **PrI**vacy and **C**ompliance Cloud

Enhanced PrIvacy and Compliance Cloud is an ISO certified cloud platform	A region of INFN Cloud with a certified Information Security Management System
EPIC Cloud offers an IaaS Community Cloud for the communities of	Biomedical and genomic researchers Industrial researchers

Site locations: Bologna (active now), Bari and Catania sites will be added in June 2024 enabling for high availability and disaster recovery

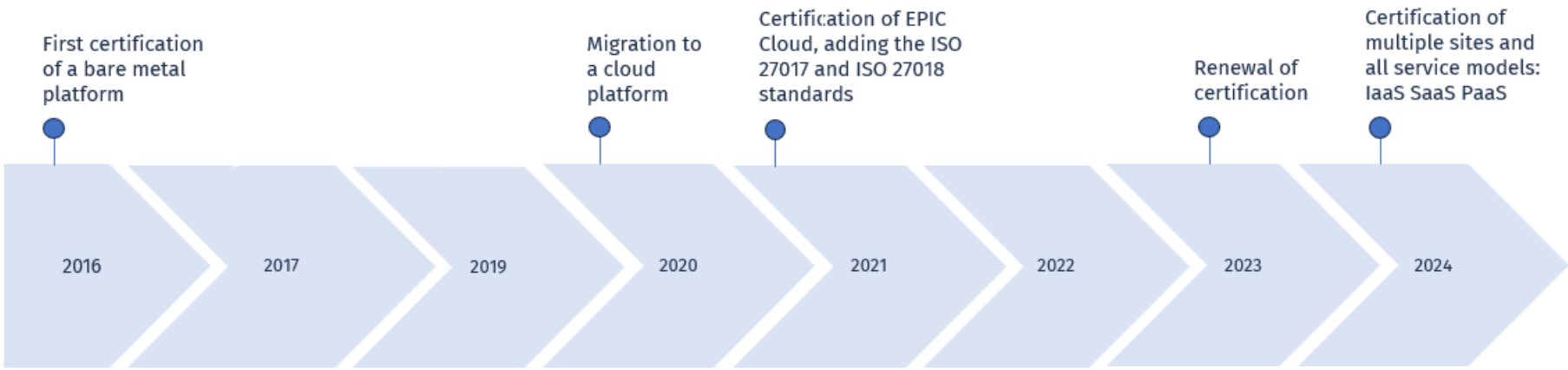
Resource available today: about 700 TB of storage, 1440 cores, 10 TB RAM, 6 GPU A100

On going expansion with 3M euro of NRRP resources and 4M euro of funds from other projects

Why EPIC

- The GDPR states that Clinical and medical data (for instance, genomic) is personal data; i.e., it fits in the Art.9 **special categories of personal data**.
  - Genomic data is mostly impossible to be anonymized → GDPR shall always be applied
- To comply with the requirements of health research projects INFN is involved in, we created a **portion of the INFN Cloud infrastructure**, applied specific organizational and technical security measures, and certified it ISO/IEC 27001, 27017, 27018.

From the Data Controller side, the fact that EPIC Cloud is ISO-certified is a way to demonstrate that processing is performed in accordance with the GDPR.

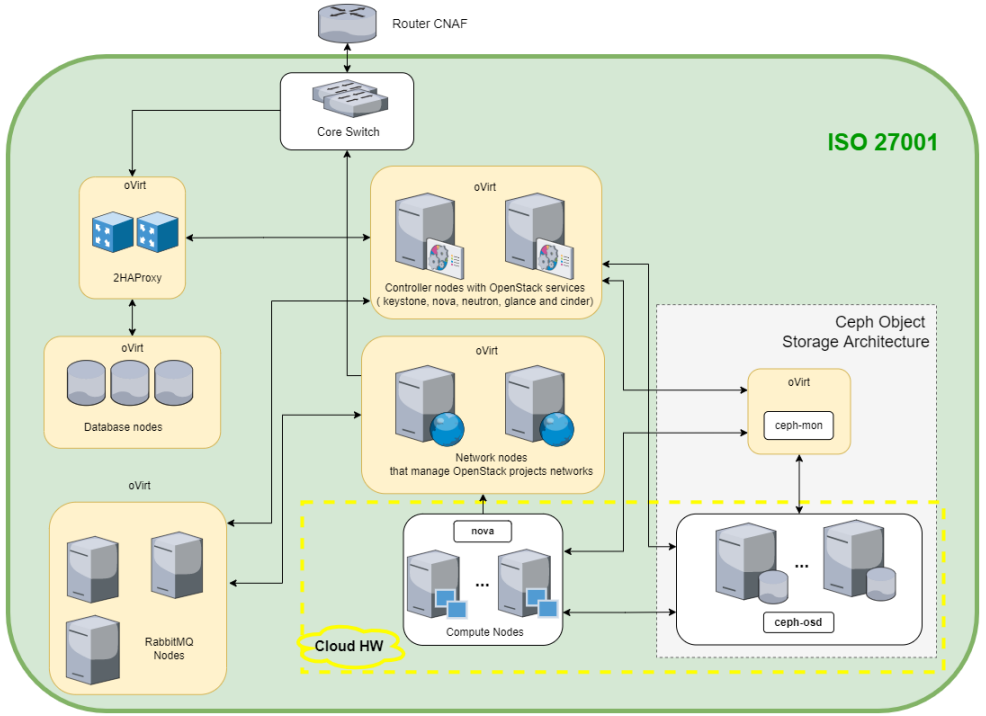


Technology

It is **based on the same technologies of INFN Cloud** (OpenStack, CEPH, IAM), with various enhancements introduced to meet higher security and privacy standards.

For example:

- OIDC with **2FA**, integration with web services, SSH and VPN (OpenVPN)
- **Network segregation** among OpenStack tenants
- At-rest and in-transit encryption
- Advanced logging and auditing services
  - **centralized syslog** server managed applying the ***segregation of duties*** principle

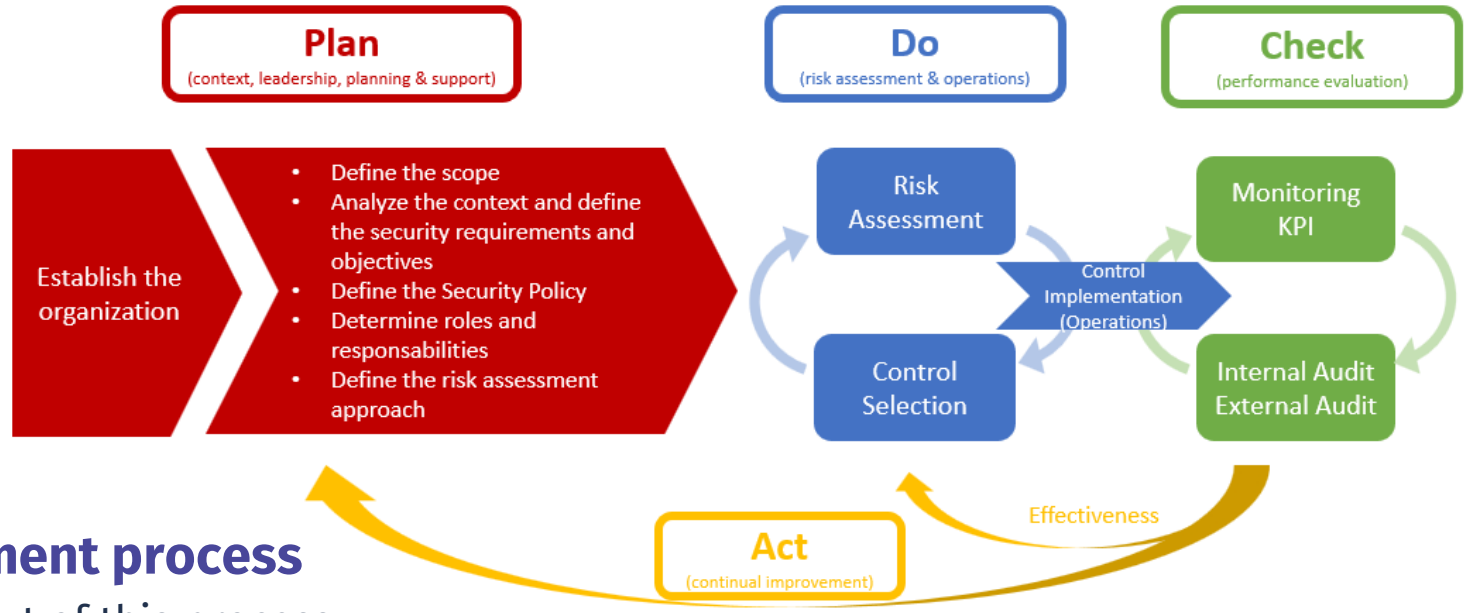


ISMS: what's all about

Information Security Management System

It is an **organizational framework** linking all the elements relevant to the information security, to assure that **policies**, **processes** and **security objectives** are implemented, communicated and assessed.

- It needs to **continually improve**
  - **Deming Cycle**



- It is centered to the **risk assessment process**
  - all decisions are based on the output of this process
- **Goal**: ensuring Confidentiality of information, while still ensuring the information remains accessible to authorized persons and is not altered

Use case examples

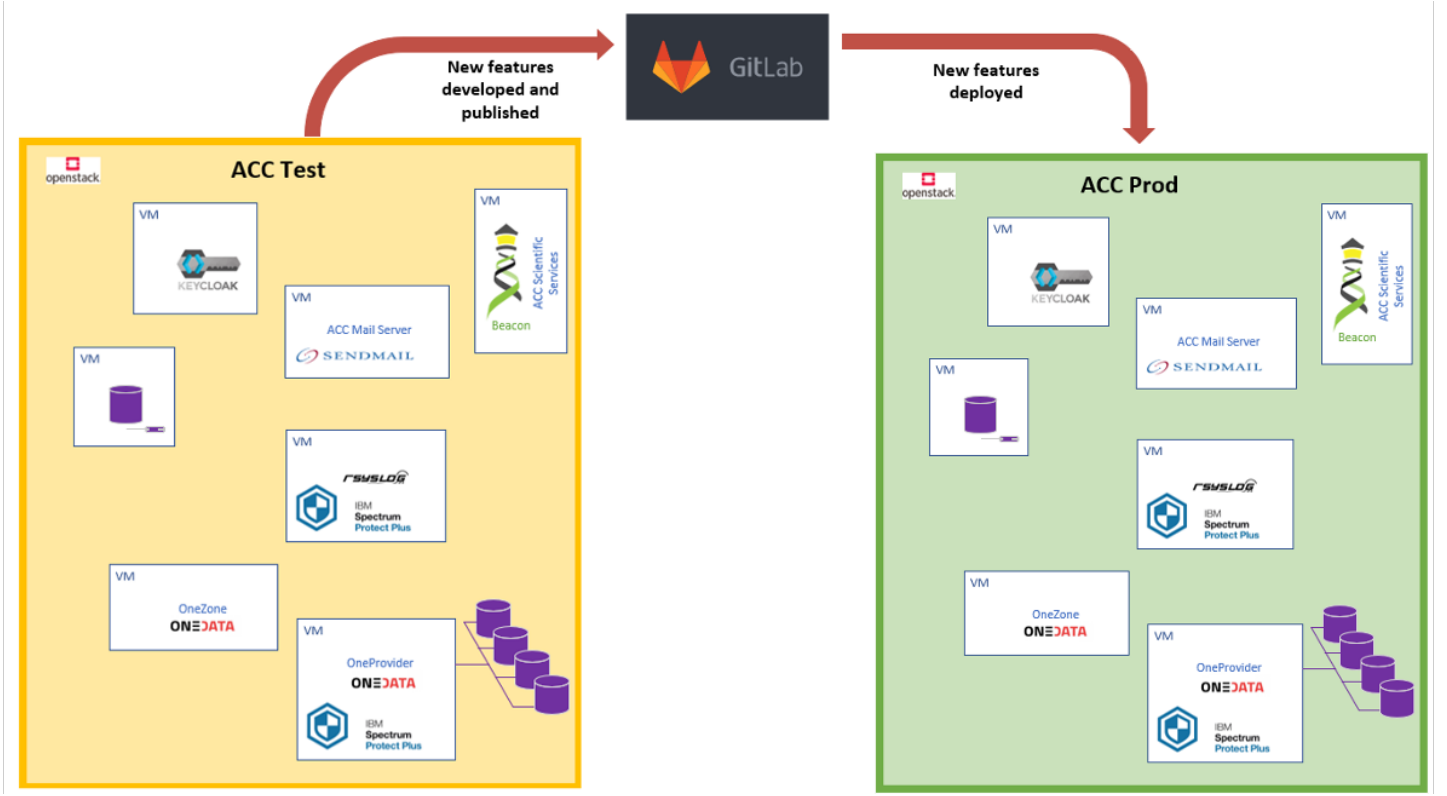
Alleanza Contro il Cancro - ACC

The National Oncology Network founded in 2002 by the Ministry of Health, joined by 51 IRCCS, ISS, AIFA, INFN and Politecnico di Milano and several patients' associations to perform translational research in the field of cancer research.

Two separate OpenStack projects:

- **ACC-Test**: services have been configured and tested and every change in configurations has been validated
- **ACC**: where services have been configured and tested and every change in configurations has been validated

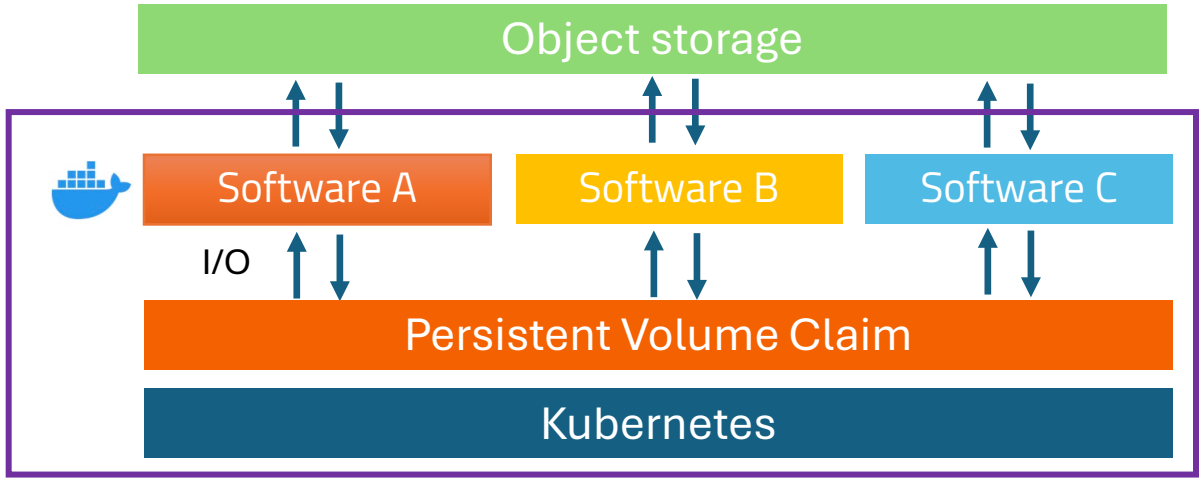
Each VM is hardened according to ISO 27001 OpenSCAP profile



INFN - IRCCS Sant'Orsola Collaboration

Joint research agreement with the following objectives

- secure applications for genomic data
- GPU -based solutions for genomic analysis methods
- federated and integrated cloud platforms for homics data
- adaptation of genomic pipelines to cloud and data lake architectures based on microservices
- Integration of omics data and other clinical data like Electronic Medical Records (EMR)



EPIC - ISO 27001

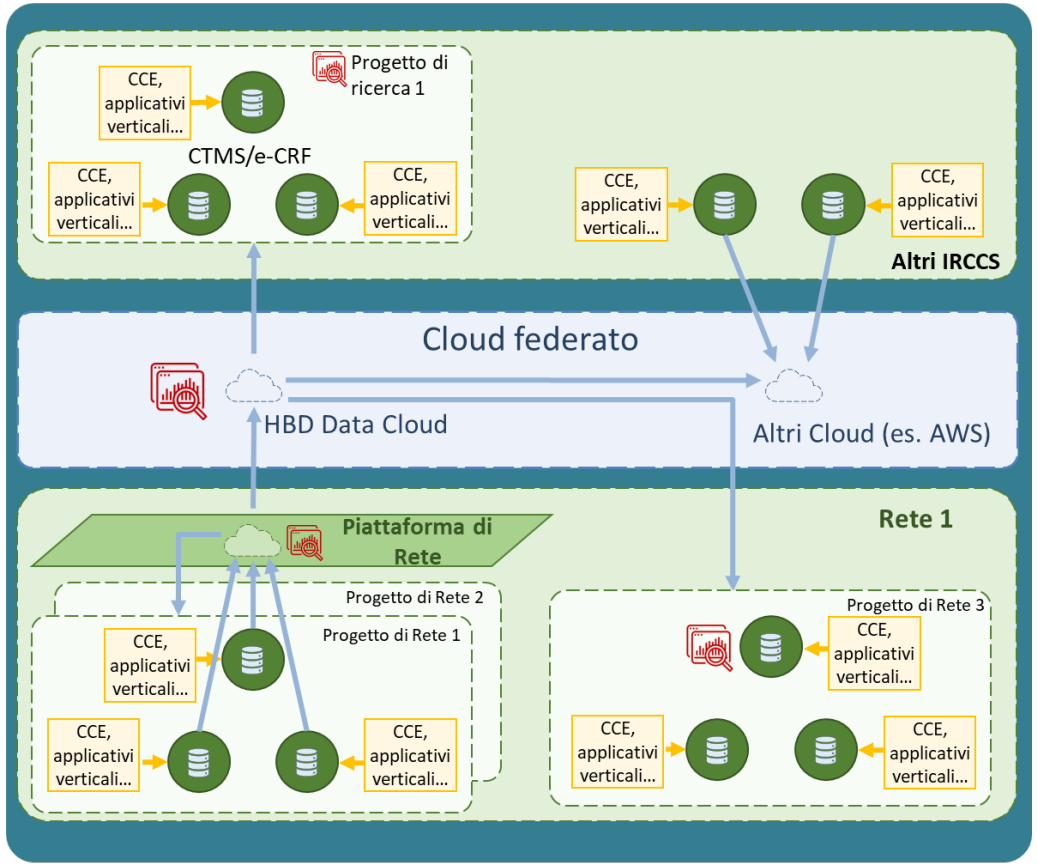
- IaaS **OpenStack** (EPIC Cloud) with *hardened OS*
- Cluster **Kubernetes**/RKE2 for orcehstration
- **Nextflow** (workflow manager)
- Monitoring with **Prometheus**/**Grafana**
- Pipeline di Continuous Integration/Continuous Delivery **(CI/CD)**

Health Big Data (HBD)

- Health Big Data is a 10-years project funded by the Italian Ministry of Health aiming at the creation of a federated and integrated big data platform for the health research at national level
  - 4 research networks: ACC, RIN, Cardio, IDEA
  - Research objectives: preventing diseases, personalizing treatments, improving the quality of life of patients
  - Budget: 55M€

Health Big Data (HBD)

- INFN is in the managing board of HBD. Its tasks include the definition of an integrated national platform and contributions to several Work Packages.
- The HBD architecture will provide solutions for several scenarios:
  1. Central harvesting of data collected remotely
  2. Edge anonymization, followed by central ingestion and analysis of data
  3. Edge feature extraction, followed by central ingestion and analysis of features
  4. Federated learning based on edge-based training, followed by publishing of the trained methods and by inference performed either centrally or at other edge locations



Contributors

Ahmad Alkhansa, Alessandro Costantini, Andrea Chierici, Cristina Vistoli, Daniele Cesini, Daniele Spiga, Diego Michelotto, Domingo Ranieri, Francesco Sinisi, Giacinto Donvito, Giusy Sergi, Jacopo Gasparetto, Letizia Magenta, Lorenzo Chiarelli, Luca dell'Agnello, Luigi Scarponi, Barbara Martelli, Nadina Foggetti, Patrizia Belluomo, Stefano Longo, Stefano Zani, Vincenzo Ciaschini