CINECA

AI in Italy

(AI on HPC)

Laura Morselli – CINECA

CINECA: A GROWING CONSORTIUM

SUPPORTING THE ITALIAN ACADEMIC SYSTEM SINCE 1969

CINECA

117 MEMBERS

Ministry of University-Research and Ministry of Education, Universities, Public Research Organisations

5 LOCATIONS

Bologna, Milano, Roma, Napoli, Chieti

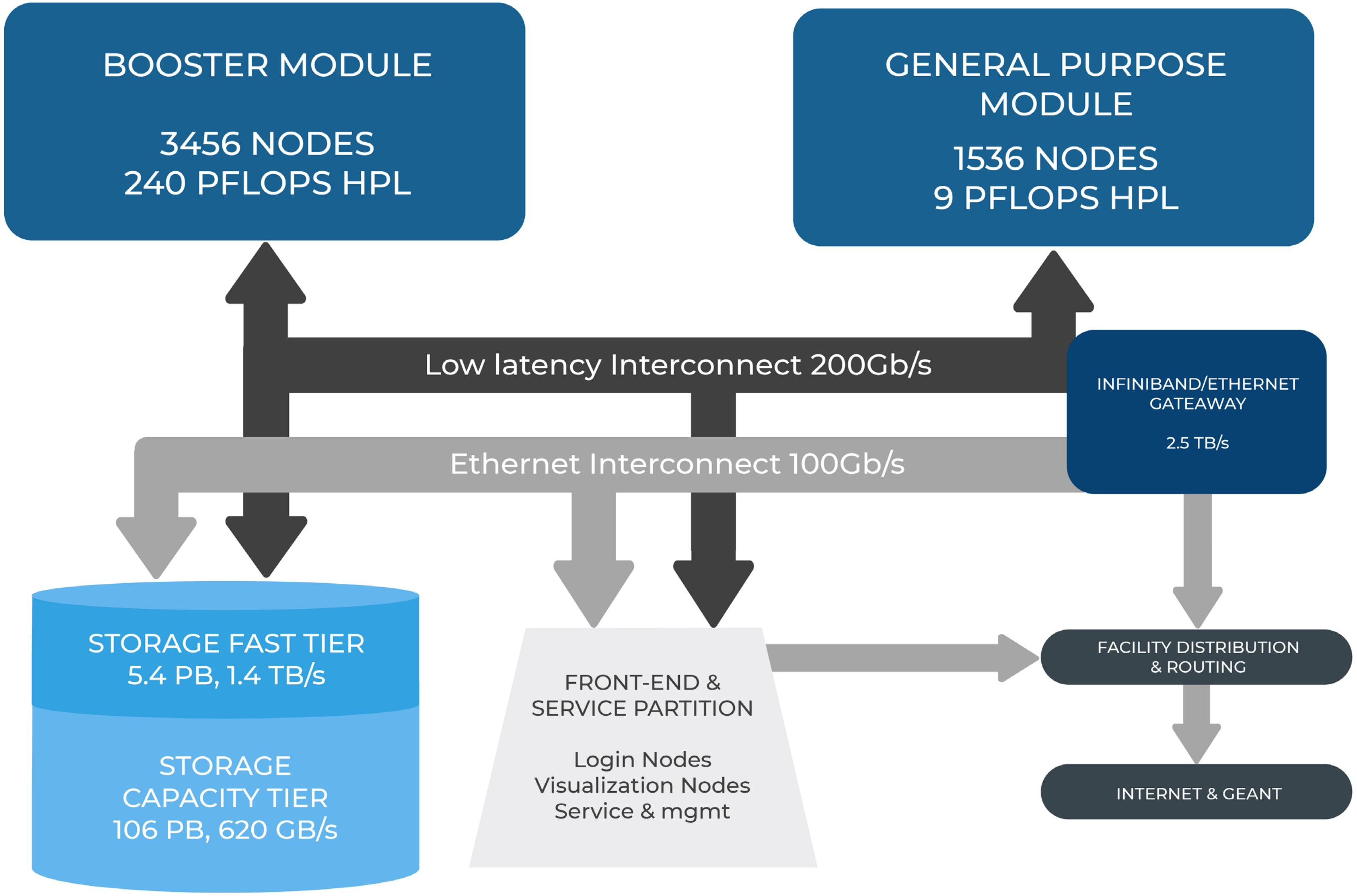
≃1100

Employees

Milano
Bologna
Chieti
Roma
Napoli

Leonardo

THE 7th MOST POWERFUL SUPERCOMPUTER IN THE WORLD



BOOSTER MODULE

3456 NODES
240 PFLOPS HPL

GENERAL PURPOSE MODULE

1536 NODES
9 PFLOPS HPL

Low latency Interconnect 200Gb/s

Ethernet Interconnect 100Gb/s

INFINIBAND/ETHERNET GATEAWAY

2.5 TB/s

STORAGE FAST TIER
5.4 PB, 1.4 TB/s

STORAGE CAPACITY TIER
106 PB, 620 GB/s

FRONT-END & SERVICE PARTITION

Login Nodes
Visualization Nodes
Service & mgmt

FACILITY DISTRIBUTION & ROUTING

INTERNET & GEANT

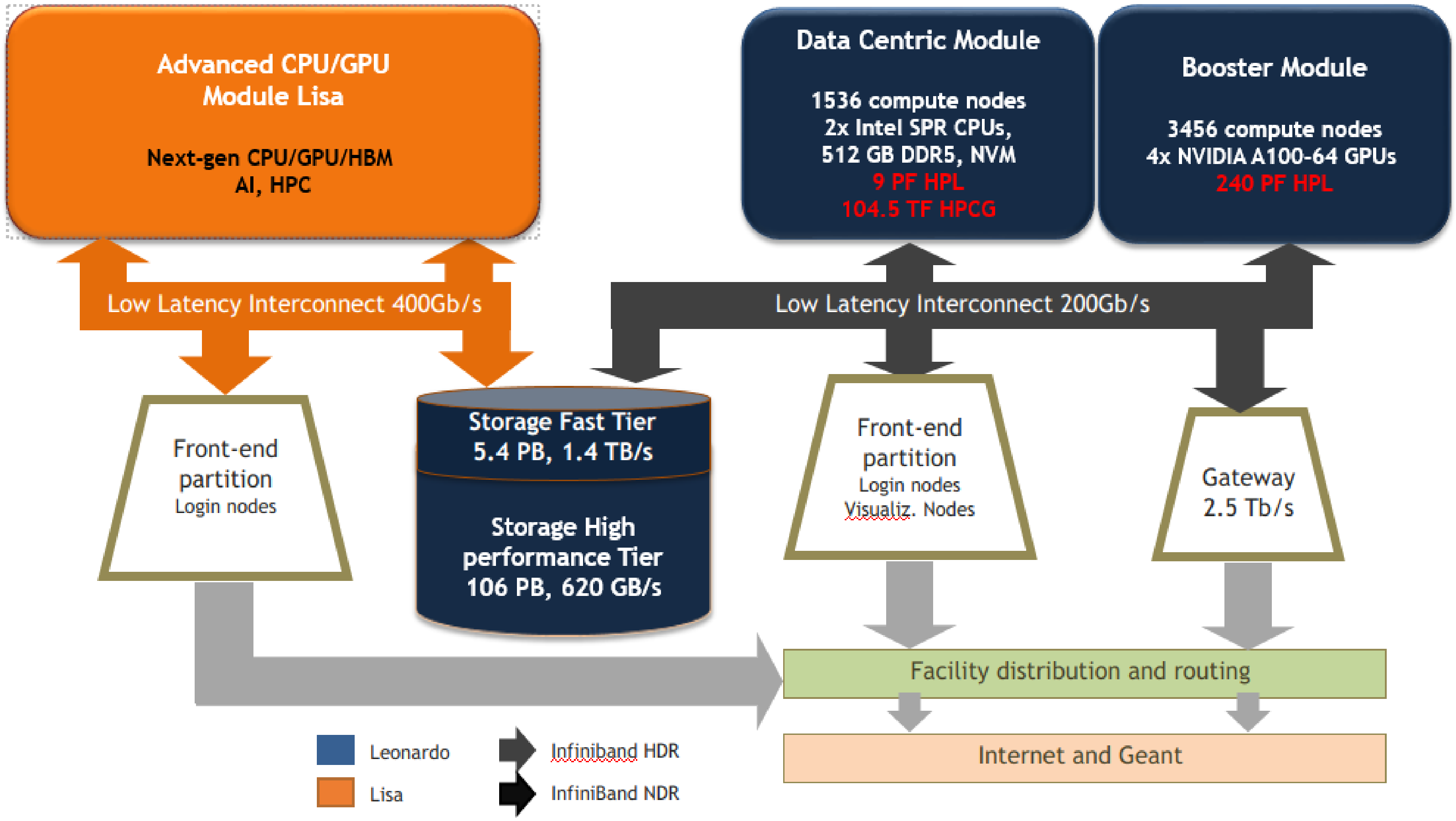
LISA

LEONARDO IMPROVED SUPERCOMPUTING ARCHITECTURE

An AI partition of Leonardo dedicated for Generative AI (technology under investigation)

3 years
65% Italy
35% EuroHPC JU

Timeframe
Q4-2024



Advanced CPU/GPU Module Lisa

Next-gen CPU/GPU/HBM
AI, HPC

Data Centric Module

1536 compute nodes
2x Intel SPR CPUs,
512 GB DDR5, NVM
9 PF HPL
104.5 TF HPCG

Booster Module

3456 compute nodes
4x NVIDIA A100-64 GPUs
240 PF HPL

Low Latency Interconnect 400Gb/s

Low Latency Interconnect 200Gb/s

Front-end partition
Login nodes

Storage Fast Tier
5.4 PB, 1.4 TB/s

Storage High performance Tier
106 PB, 620 GB/s

Front-end partition
Login nodes
Visualiz. Nodes

Gateway
2.5 Tb/s

Facility distribution and routing

Internet and Geant

🟦 Leonardo
🟧 Lisa

Infiniband HDR
InfiniBand NDR

Hundreds of GBs of shared GPU memory for Large AI Models!

ISCRA

CALL FOR PROPOSALS OF THE ITALIAN SUPERCOMPUTING RESOURCE ALLOCATION ORGANIZATION

- **Class B (Max 250'000 GPU hours on Leonardo Booster – 12 months)**
  Class B projects are received twice a year. They go under peer-review evaluation and a 5 months period is expected before access to HPC resources.

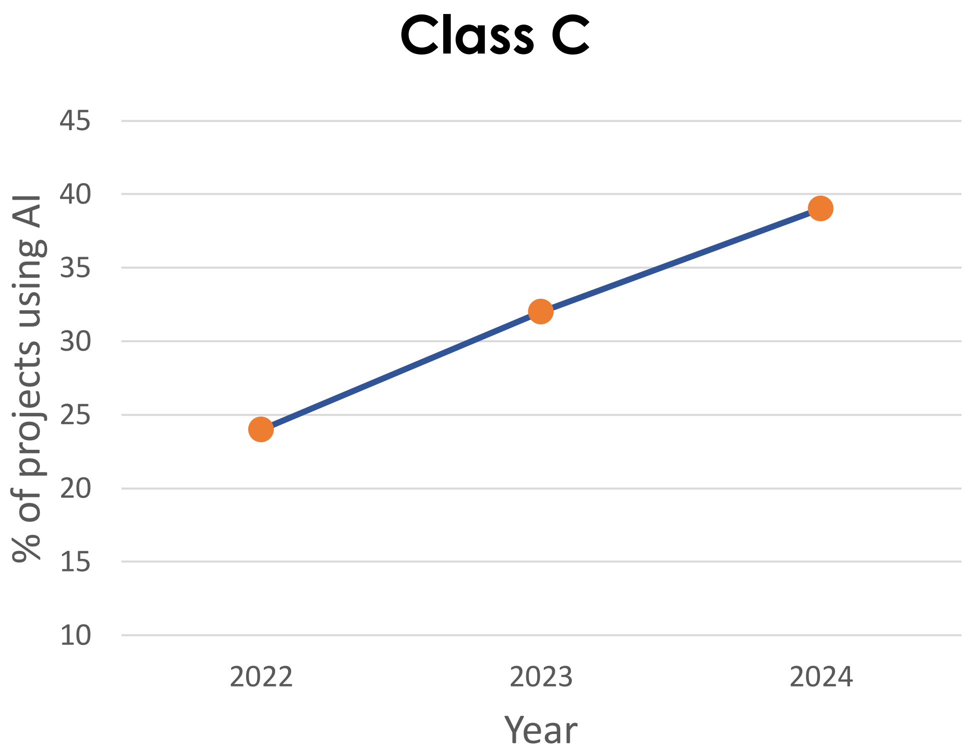
- **Class C (Max 10'000 GPU hours on Leonardo Booster – 9 months)**
  Class C projects are received through continuous submission and reviewed once per month. An average period of about 30 days is required for activating the project.

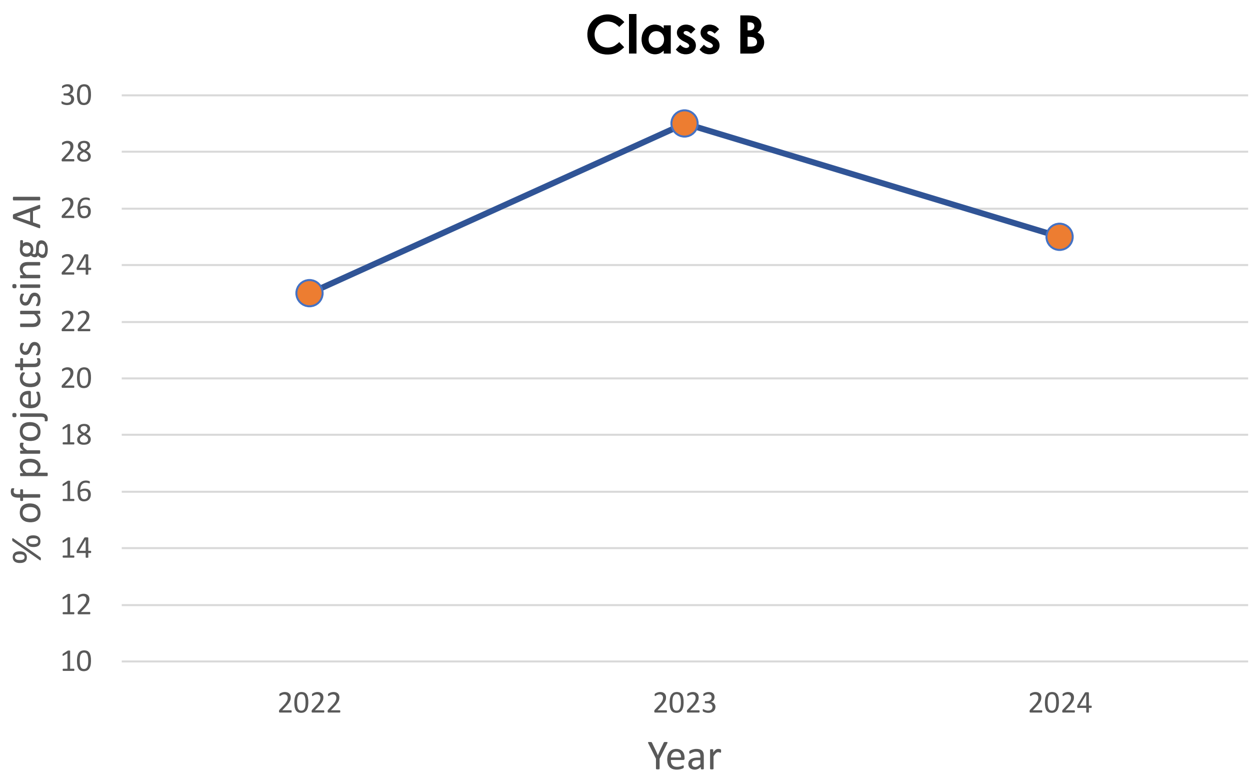
Projects' PIs need to be affiliated to an Italian research institution, while no restriction is applied for the Co-PI and collaborators. It is expected that the research will be performed at Italian institutions.

CINECA provides **high-level technical support** to each project through its User Support Group. Dedicated specialist support can be requested for the enabling and optimization of the applications necessary for the project

ISCRA

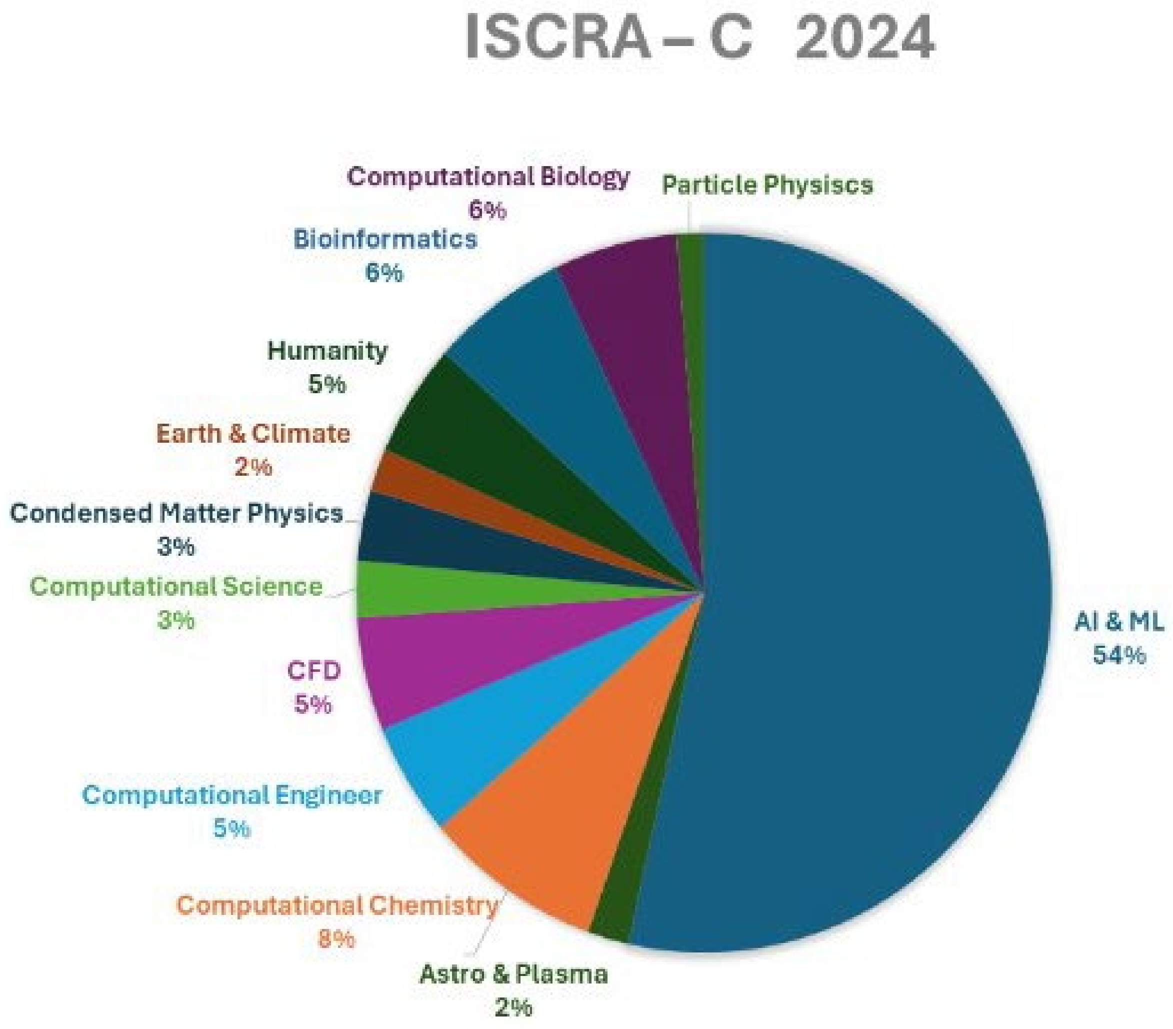
CALL FOR PROPOSALS OF THE ITALIAN SUPERCOMPUTING RESOURCE ALLOCATION ORGANIZATION



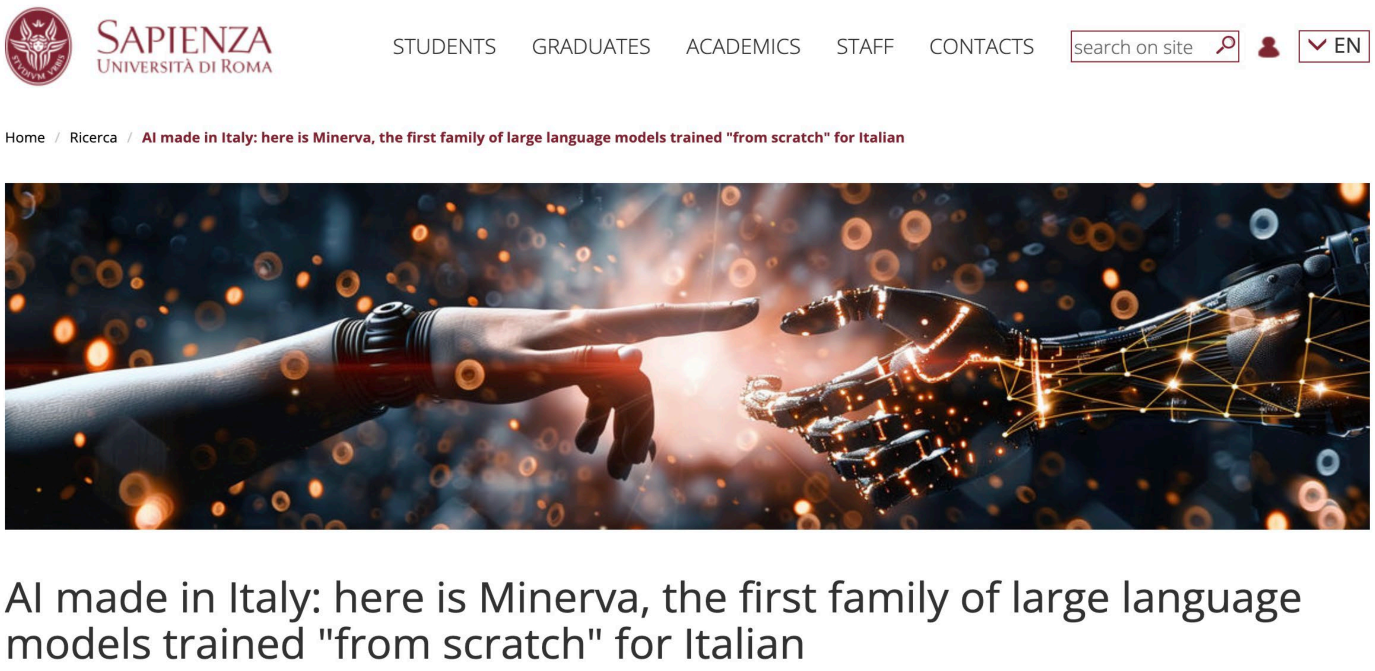


ISCRA

CALL FOR PROPOSALS OF THE ITALIAN SUPERCOMPUTING RESOURCE ALLOCATION ORGANIZATION



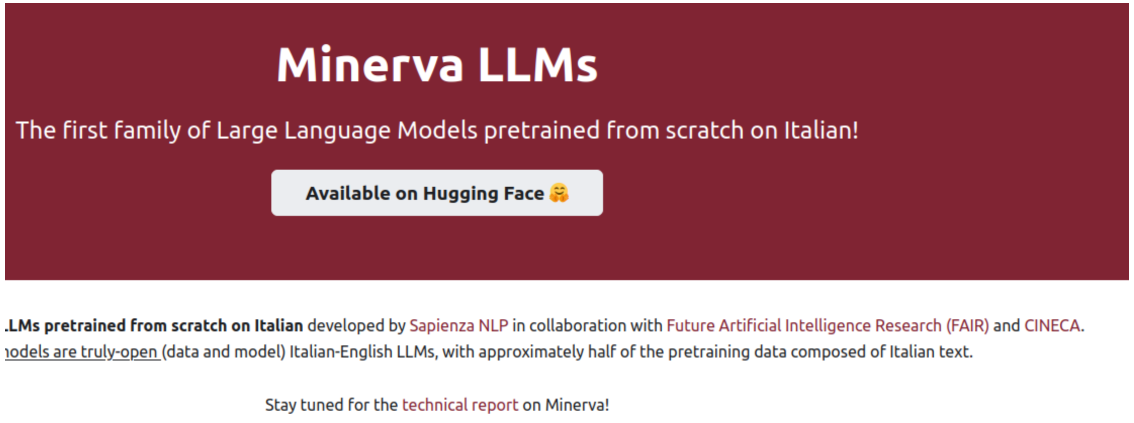
LLMs – LMMs related projects



Home / Ricerca / AI made in Italy: here is Minerva, the first family of large language models trained "from scratch" for Italian

AI made in Italy: here is Minerva, the first family of large language models trained "from scratch" for Italian

Future Artificial Intelligence Research

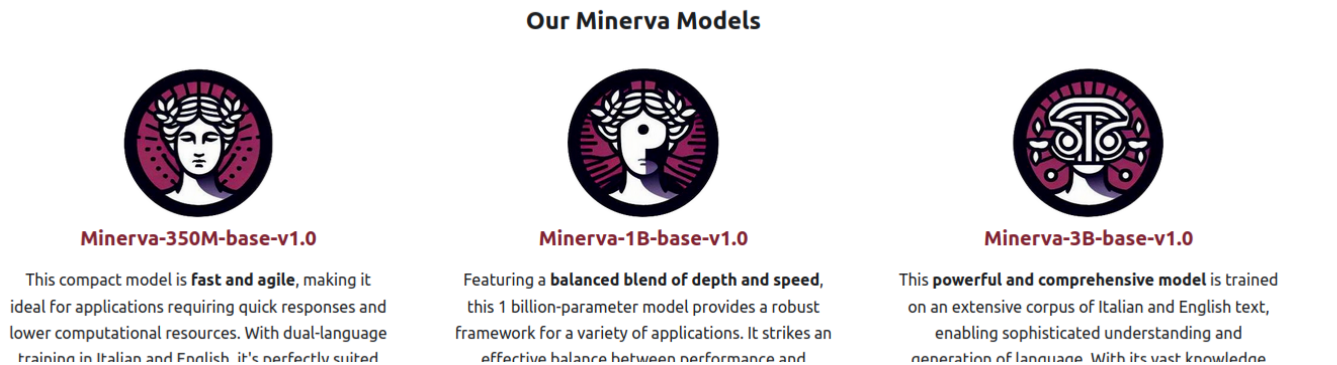


Minerva LLMs

The first family of Large Language Models pretrained from scratch on Italian!

.LMs **pretrained from scratch on Italian** developed by Sapienza NLP in collaboration with Future Artificial Intelligence Research (FAIR) and CINECA.
ıodels are truly-open (data and model) Italian-English LLMs, with approximately half of the pretraining data composed of Italian text.

Stay tuned for the technical report on Minerva!



Our Minerva Models

Minerva-350M-base-v1.0

This compact model is **fast and agile**, making it ideal for applications requiring quick responses and lower computational resources. With dual-language training in Italian and English, it's perfectly suited

Minerva-1B-base-v1.0

Featuring a **balanced blend of depth and speed**, this 1 billion-parameter model provides a robust framework for a variety of applications. It strikes an effective balance between performance and

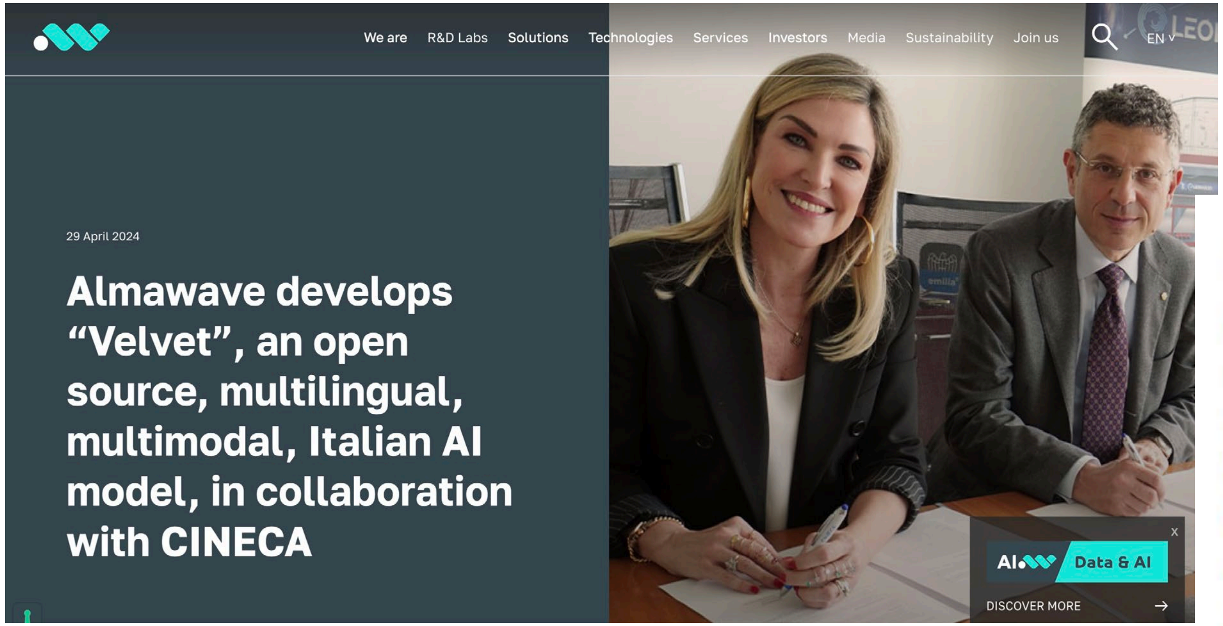
Minerva-3B-base-v1.0

This **powerful and comprehensive model** is trained on an extensive corpus of Italian and English text, enabling sophisticated understanding and generation of language. With its vast knowledge

LLMs – LMMs related projects

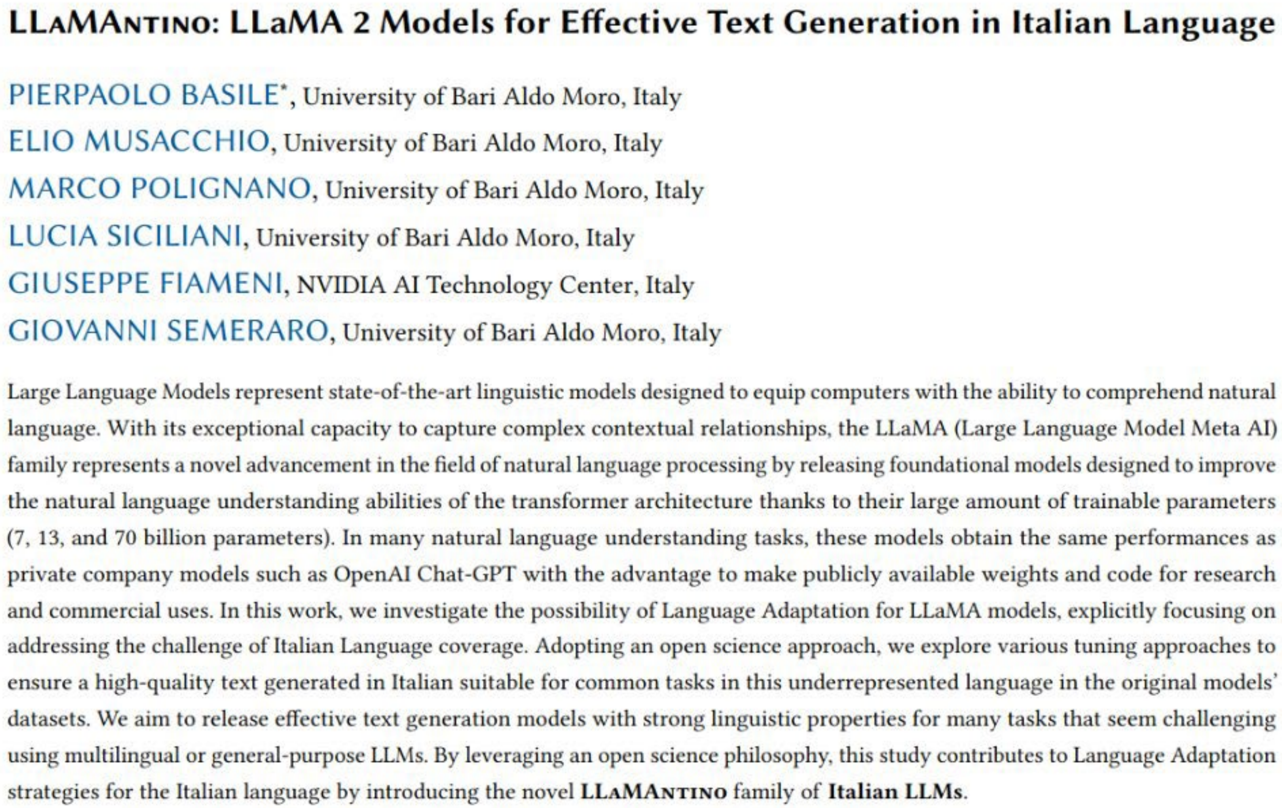
iGenius and Cineca partner for Modello Italia, the first Italian Foundational Large Language Model

LLMs – LMMs related projects



29 April 2024

Almawave develops "Velvet", an open source, multilingual, multimodal, Italian AI model, in collaboration with CINECA



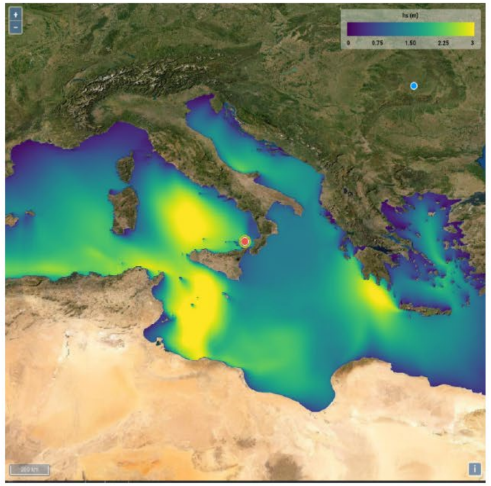
LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language

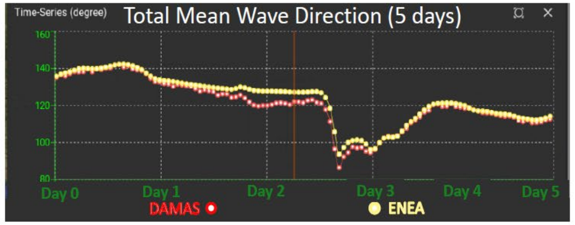
PIERPAOLO BASILE*, University of Bari Aldo Moro, Italy
ELIO MUSACCHIO, University of Bari Aldo Moro, Italy
MARCO POLIGNANO, University of Bari Aldo Moro, Italy
LUCIA SICILIANI, University of Bari Aldo Moro, Italy
GIUSEPPE FIAMENI, NVIDIA AI Technology Center, Italy
GIOVANNI SEMERARO, University of Bari Aldo Moro, Italy

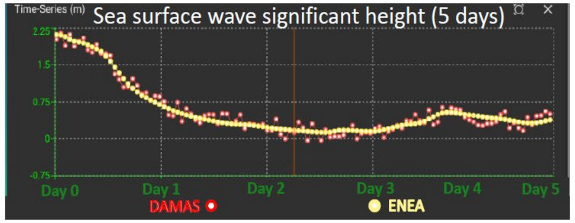
Large Language Models represent state-of-the-art linguistic models designed to equip computers with the ability to comprehend natural language. With its exceptional capacity to capture complex contextual relationships, the LLaMA (Large Language Model Meta AI) family represents a novel advancement in the field of natural language processing by releasing foundational models designed to improve the natural language understanding abilities of the transformer architecture thanks to their large amount of trainable parameters (7, 13, and 70 billion parameters). In many natural language understanding tasks, these models obtain the same performances as private company models such as OpenAI Chat-GPT with the advantage to make publicly available weights and code for research and commercial uses. In this work, we investigate the possibility of Language Adaptation for LLaMA models, explicitly focusing on addressing the challenge of Italian Language coverage. Adopting an open science approach, we explore various tuning approaches to ensure a high-quality text generated in Italian suitable for common tasks in this underrepresented language in the original models' datasets. We aim to release effective text generation models with strong linguistic properties for many tasks that seem challenging using multilingual or general-purpose LLMs. By leveraging an open science philosophy, this study contributes to Language Adaptation strategies for the Italian language by introducing the novel **LLaMAntino** family of **Italian LLMs**.

DAMAS

DATA-DRIVEN MODEL FOR THE ANALYSIS OF THE SEA STATE





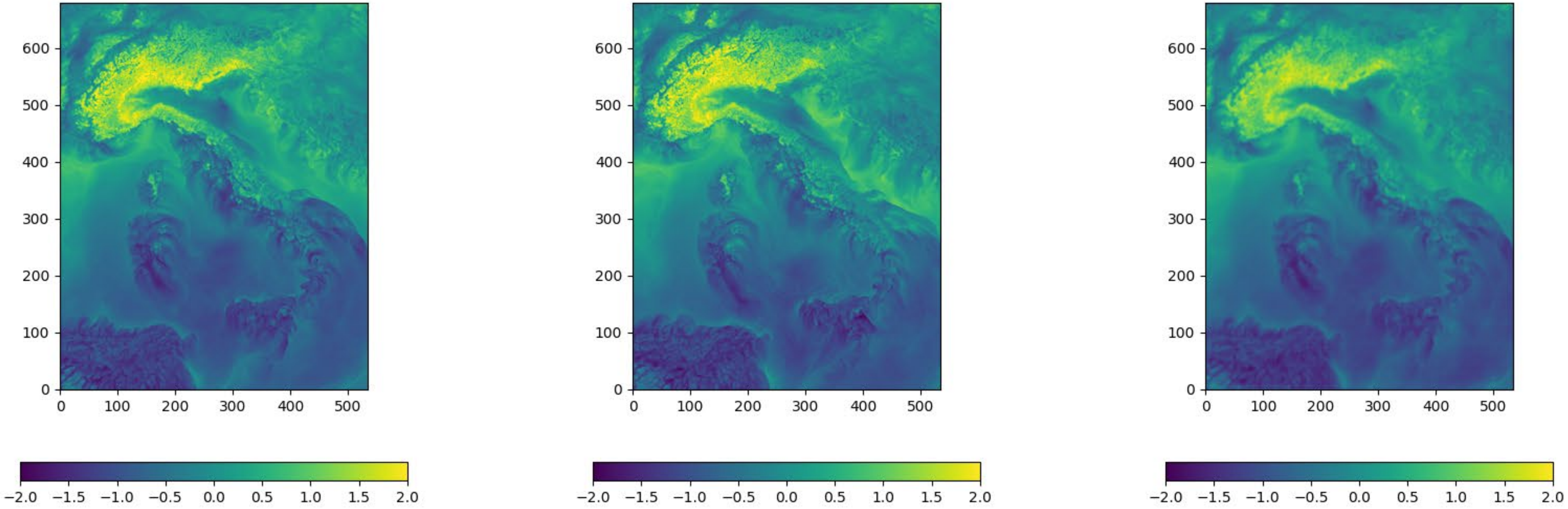


Gmatics AI models engineered on CINECA HPC infrastructure reduced the execution time and increased the resolution and the accuracy of sea circulation models and height of waves prediction, providing as a result:

- AI models that replicate the ENEA MITO and WAVE physical forecast models and produce **forecasts** over the next 5 days with hourly temporal resolution,
- The **nowcast** chains for wave and circulation for the next 12 hours,
- **Hindcast** pipelines for the historical analysis of wave and circulation sea parameters and
- a graphical user environment

AI-GCM

AI GENERAL CIRCULATION MODEL



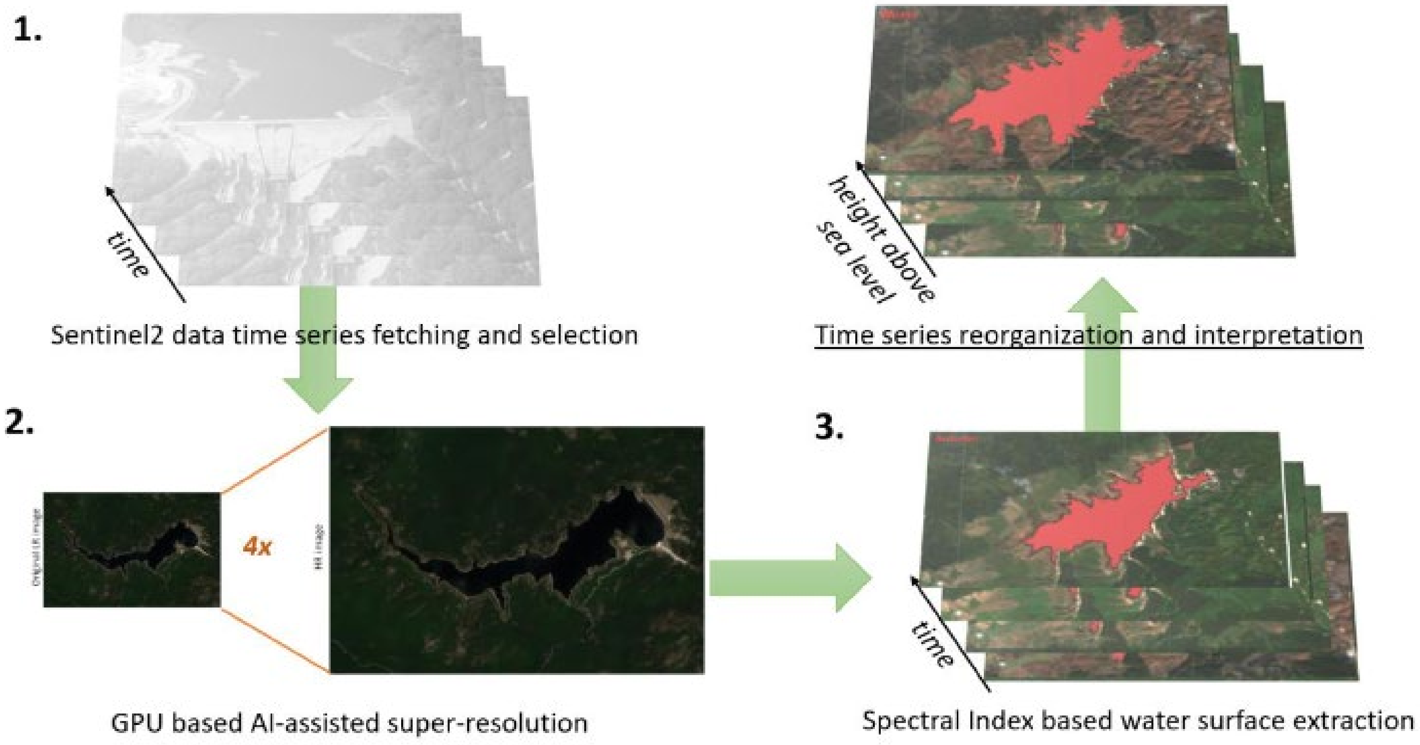
The project aims to construct a **novel GCM approach** using recent developments in ML and supported by intra-seasonal atmospheric signals. The aim is twofold. On the one hand, the potential to **significantly reduce the run time of short-term forecasts.** On the other hand, the possibility to **increase the forecast skill for the medium term.** The final objective is to produce a prototype model that could be further developed with future funding.

HPC4RM

HPC FOR RESERVOIR MONITORING

- **Continuous monitoring of water reservoirs is gaining importance** due to water shortages, periods of drought, and extreme weather events in all European countries.

- The amount of data and the required processing sequence calls for **on-demand HPC resources** in order to deliver necessary added value to customers.



Sentinel2 data time series fetching and selection

GPU based AI-assisted super-resolution

Spectral Index based water surface extraction

Time series reorganization and interpretation

HPC4RM

HPC FOR RESERVOIR MONITORING

THE SOLUTION

- Aresys designed a **new service:** S2 optical images are used to identify **reservoir water surface variation over time**, without the need for in-situ measurements.

- The water surface is estimated based on proper geometric and radiometric processing using the latest Deep Learning results to cope with the Sentinel-2 data resolution.

- HPC resources are necessary for the AI-assisted processing steps and to provide the starting 3D reservoir model based on historical data sets.

WEATHER AI

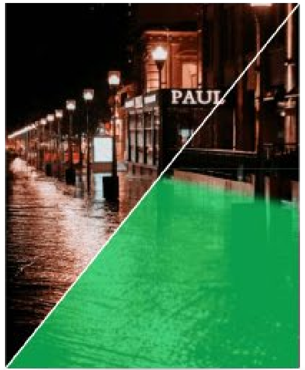
WEATHER NOWCASTING

EUROCC ITALY
HPC Competence Center

waterview
your eye on climate change

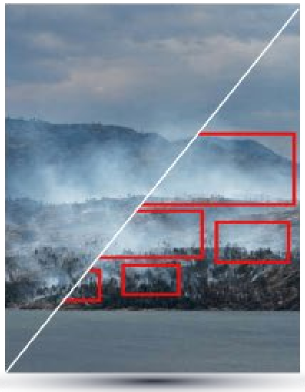
"**AI technologies** to maximize cameras potential to derive **actionable weather and environmental insights** in support of your decision making"

floodcam

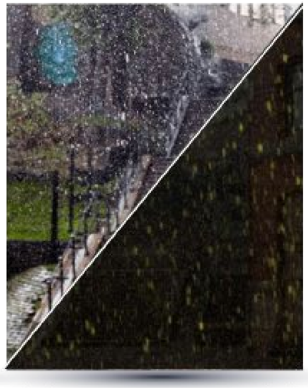


snowcam

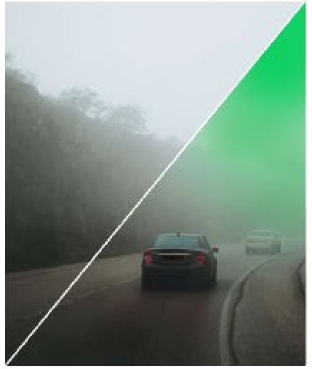
smocam



weathercam



viscam



WEATHER AI

WEATHER NOWCASTING

THE SOLUTION

- Distributed training of NN --> increased accuracy thanks to larger datasets
- Parallel Image Augmentation on CPUs
- Embarassingly parallel hyperparameter optimization

CINECA

Thank you!

Laura Morselli – CINECA