Iniziativa per l'accesso delle start-up etiche e responsabili del settore dell'intelligenza artificiale alle capacità dei supercomputer europei

Gabriella Scipione

LUNEDÌ 19 FEBBRAIO 2024 - ORE 09:45
presso Sala B-C - Viale Aldo Moro 50,
Bologna

CINECA

EuroHPC e CINECA

The Scientific Cases for computing in Europe 2018-2026

We are witnessing a revolution in humankind's ability to solve complex problems by relying on the synergy of advanced algorithms, data, and hardware.

The US, China and Japan are making great strides in these frontiers, and we call attention to the urgent need for an expanded European advanced computing infrastructure …

Simulations are critical in Climate, Weather, and Earth Sciences. Exascale resources will enable sub-kilometre resolution instead of 10km

Data is driving a scientific revolution that relies heavily on computing to process, analyse, and translate information into knowledge and technological innovations.

Computing is undergoing a tectonic change ….for hardware and extensive deployment of accelerator technologies where traditional modelling is increasingly complemented by data-driven approaches and artificial intelligence.

By the PRACE Scientific Steering Committee

EuroHPC Joint Undertaking: 34 countries + EC

#EuroHPC
Joint Undertaking

The European High Performance Computing Joint Undertaking (EuroHPC JU) will pool European resources to develop top-of-the range exascale supercomputers for processing big data, based on competitive European technology.

Member countries are Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, the Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden and Turkey.

EuroHPC
Joint Undertaking

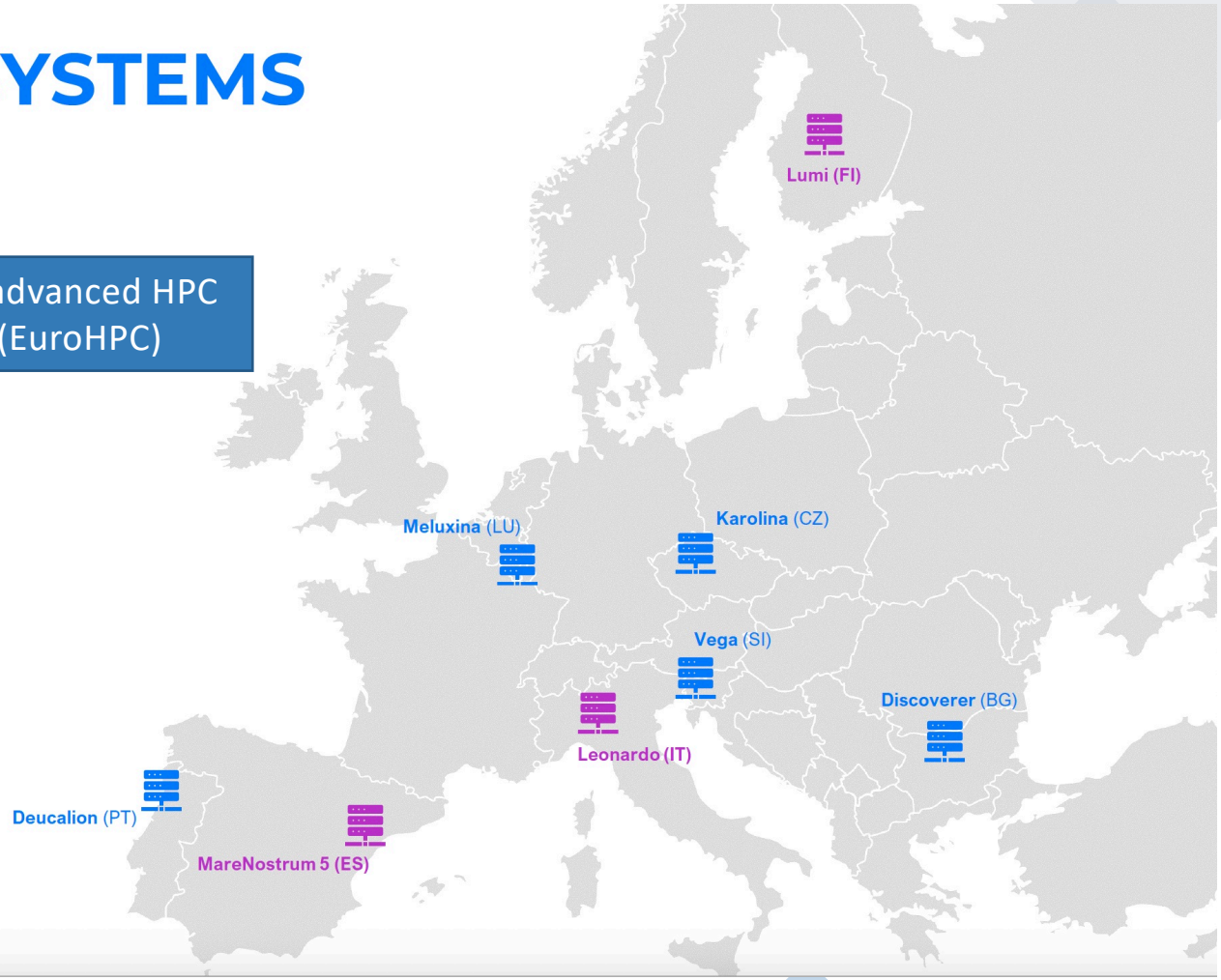
European Commission

EUROHPC SYSTEMS
2019 → 2023

World's most advanced HPC infrastructure (EuroHPC)

PRE-EXASCALE

PETASCALE



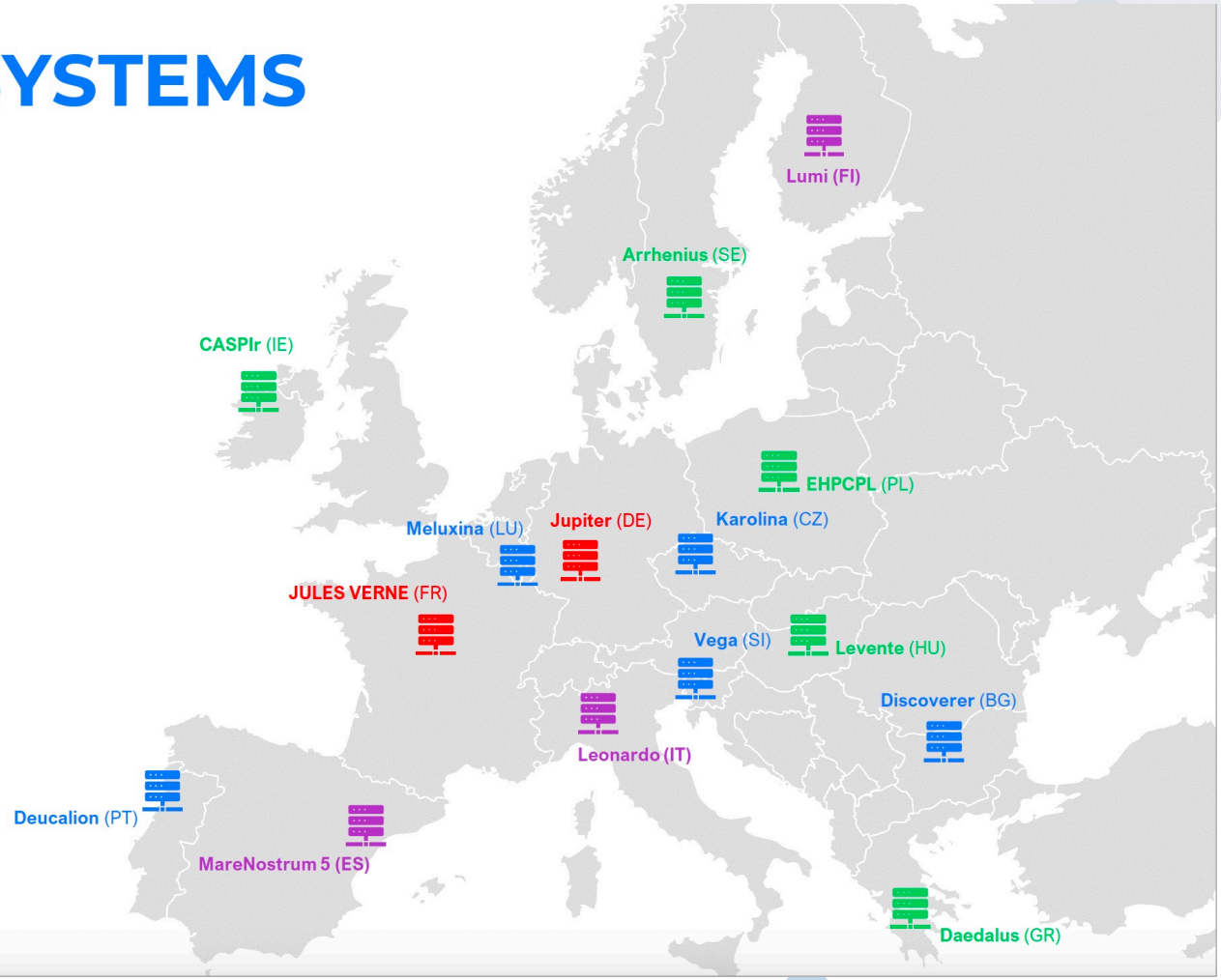
EUROHPC SYSTEMS
2024 → 2026

EXASCALE

PRE-EXASCALE

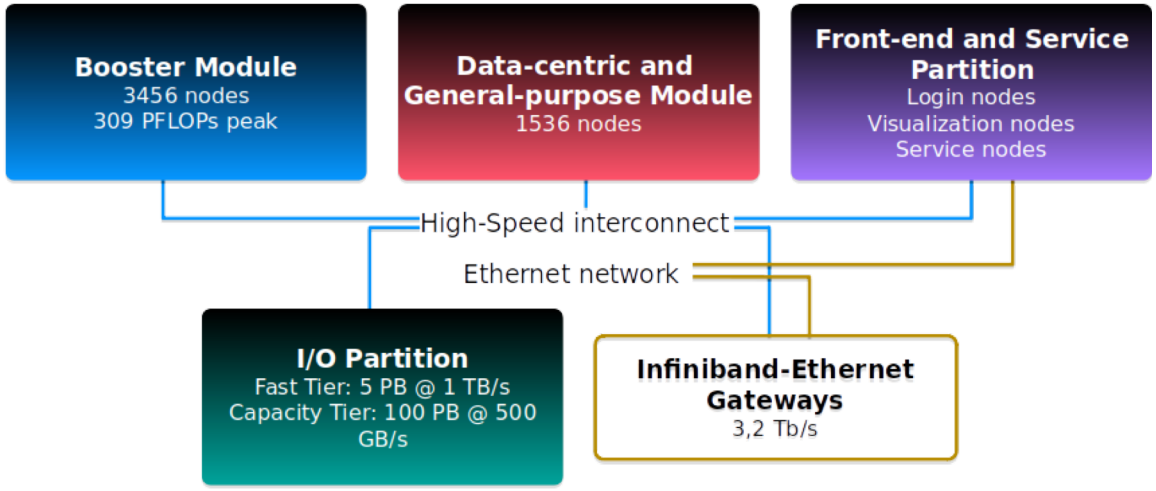
PETASCALE

MID-RANGE



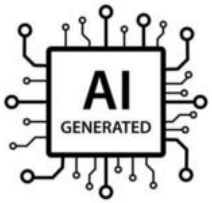
Cineca (EuroHPC) Infrastructure

General features and performance



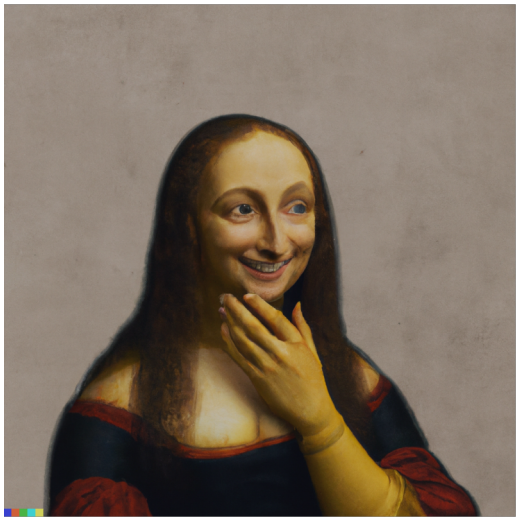
LISA

THE LEONARDO'S UPGRADE



LISA

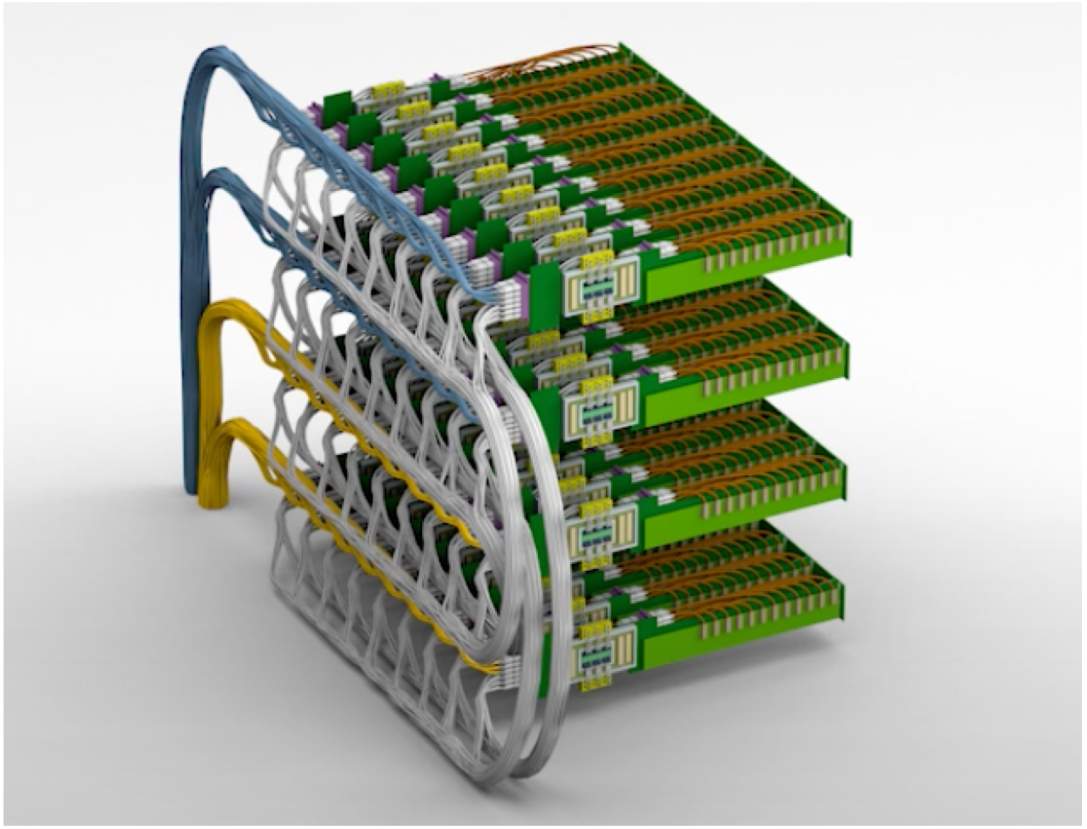
Leonardo Improved Supercomputing Architecture



Courtesy: DALL-E 2

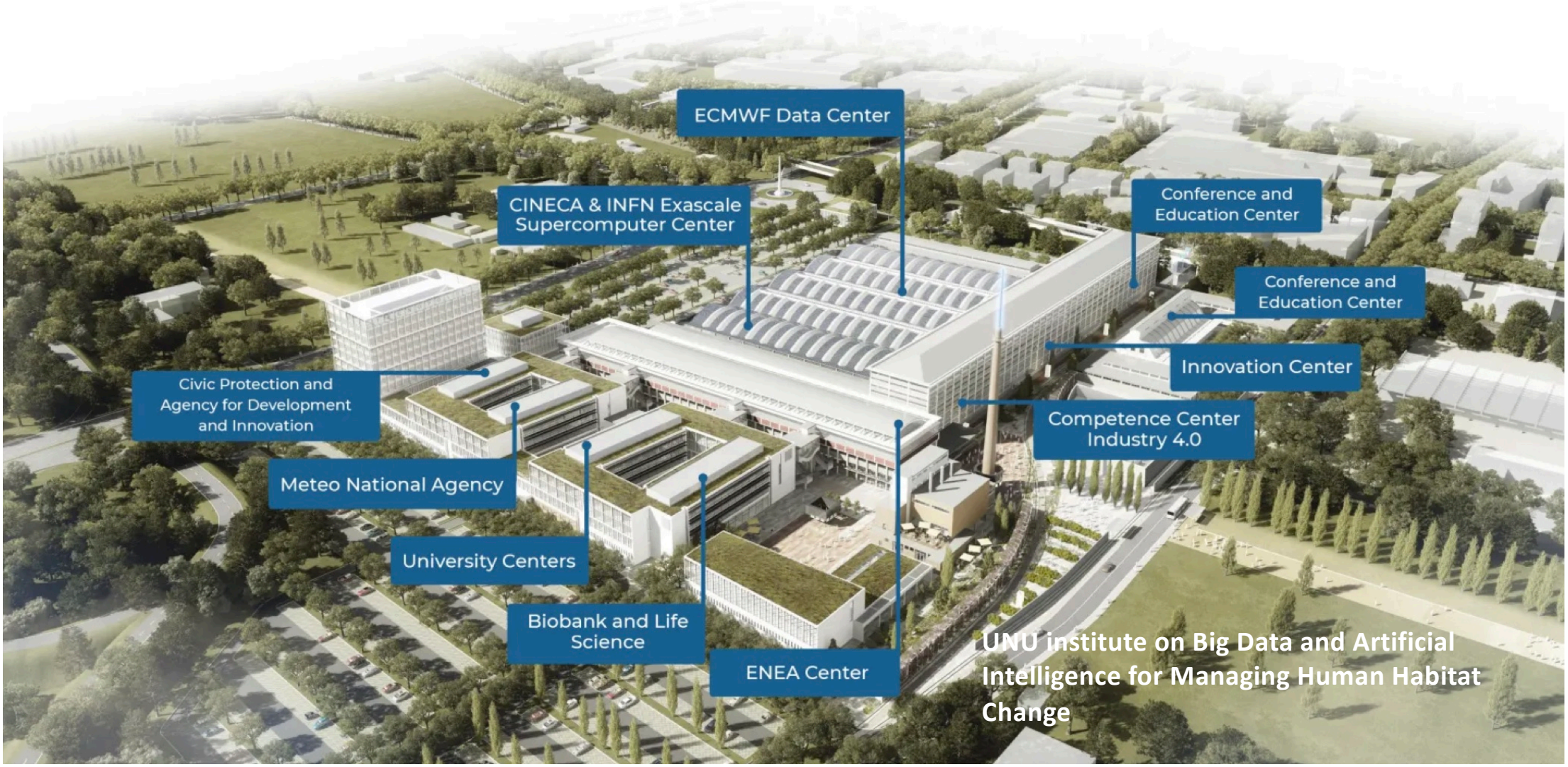
3 years
65% Italy + 35% EuroHPC JU

An AI partition of Leonardo dedicated for Generative AI



TBs of shared GPU memory for LLMs!

Tecnopolo di Bologna: Cineca Data Center



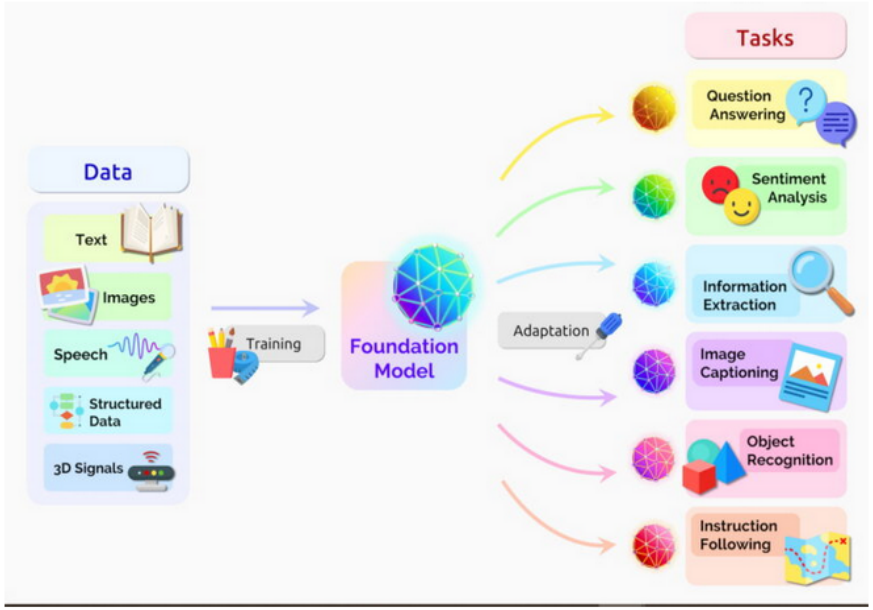
AI, LLM e HPC

Foundation models

• I **Foundation Models** sono modelli di AI addestrati su grandi insiemi di dati e possono imparare a riconoscere modelli, fare previsioni ed eseguire altri compiti senza essere esplicitamente programmati per farlo.

• Forniscono una struttura per lo **sviluppo di modelli più complessi** e possono migliorare notevolmente l'accuratezza e l'efficienza dei sistemi di intelligenza artificiale.

• Senza i foundation models, molti dei **recenti progressi** nell'AI non sarebbero stati possibili.



Source: On the Opportunities and Risks of Foundation Models

Foundation models

Modelli linguistici

Sono progettati per **comprendere e generare il linguaggio umano**. Eccellono in compiti come il completamento di testi, la traduzione, la sintesi e persino la scrittura creativa. Sono conosciuti anche come **Large Language Models** (LLM).

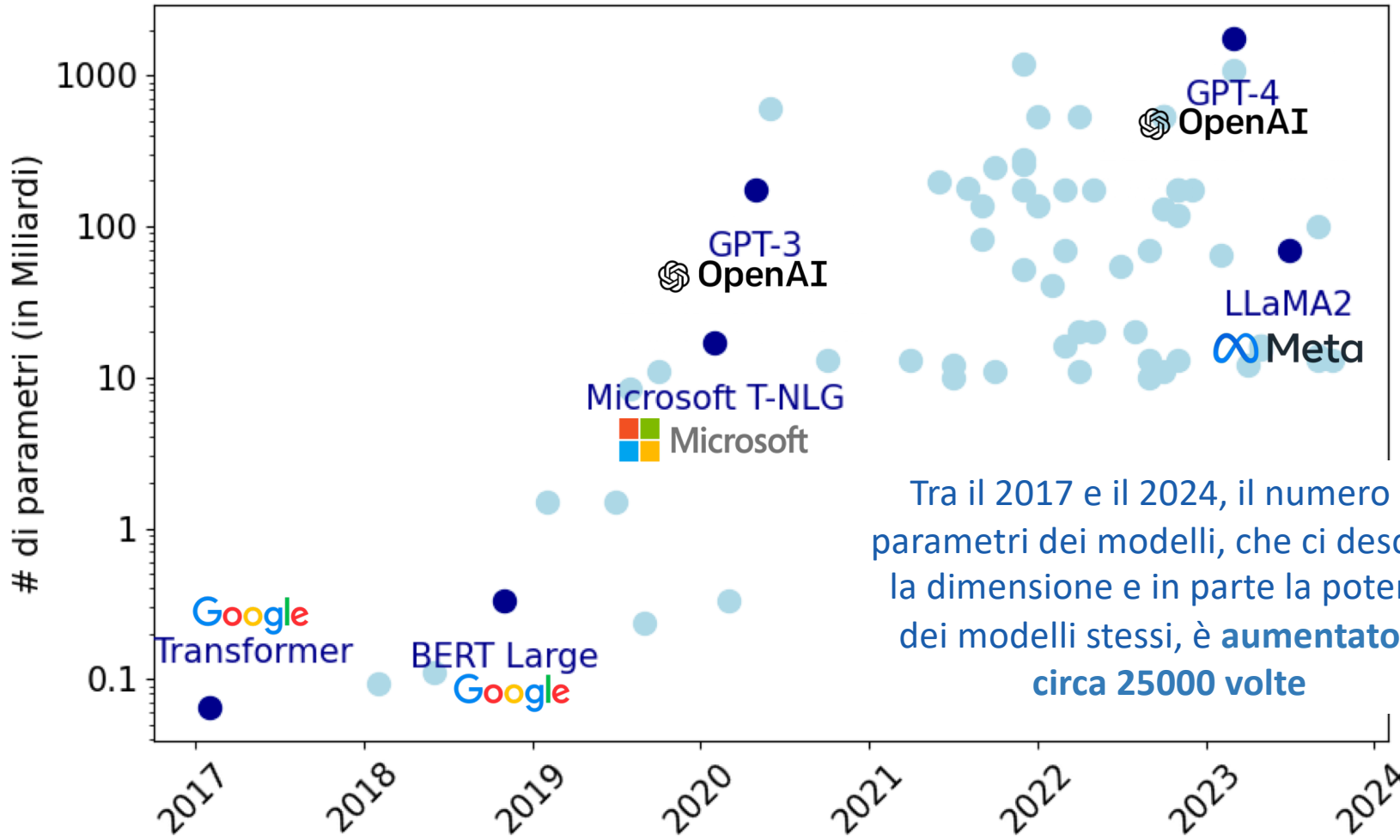
Modelli visivi

Sono specializzati in compiti di computer vision, consentendo alle macchine di comprendere e interpretare **dati visivi come immagini e video**. Hanno dimostrato prestazioni eccezionali in compiti come la classificazione delle immagini, il rilevamento degli oggetti e la generazione di immagini.

Modelli multimodali

I modelli multimodali combinano la potenza dei modelli linguistici e di visione, consentendo alle macchine di **comprendere e generare contenuti testuali, audio e visivi**. Questi modelli sono in grado di elaborare simultaneamente informazioni provenienti da modalità diverse, consentendo di fare progressi in attività come la didascalia delle immagini, la risposta a domande visive e persino la generazione di descrizioni coerenti da input di immagini o audio.

Evoluzione dei Large Language Models



Tra il 2017 e il 2024, il numero di parametri dei modelli, che ci descrive la dimensione e in parte la potenza dei modelli stessi, è **aumentato di circa 25000 volte**

Sources: A Survey of Large Language Models ; Memory Footprint and FLOPs for SOTA Models in CV/NLP/Speech

AI e High Performance Computing

Scalabilità
I task di AI spesso richiedono l'analisi di enormi quantità di dati e la manipolazione di modelli complessi, il che richiede una potenza di calcolo elevata per gestire il carico di lavoro in modo efficiente.

Velocità
L'addestramento di modelli AI può richiedere settimane o addirittura mesi utilizzando risorse di calcolo standard. Con l'HPC, è possibile ridurre drasticamente i tempi di addestramento, consentendo un ciclo di sviluppo più rapido.

Prestazioni
L'HPC consente di eseguire più esperimenti in parallelo, ottimizzando i modelli AI e migliorando le prestazioni complessive.

Il processo dei modelli AI

Training

Il training è il processo tramite il quale un modello AI impara a riconoscere pattern nei dati attraverso interazioni successive sui dati forniti. Durante questo processo, il modello ottimizza i suoi parametri minimizzando l'errore.

Fine-tuning

Il fine-tuning è il processo di adattamento di un modello AI già addestrato per eseguire compiti specifici o per migliorare le sue prestazioni su nuovi dati. Questo coinvolge l'aggiornamento dei pesi del modello utilizzando un set di dati più piccolo e specifico.

Inference

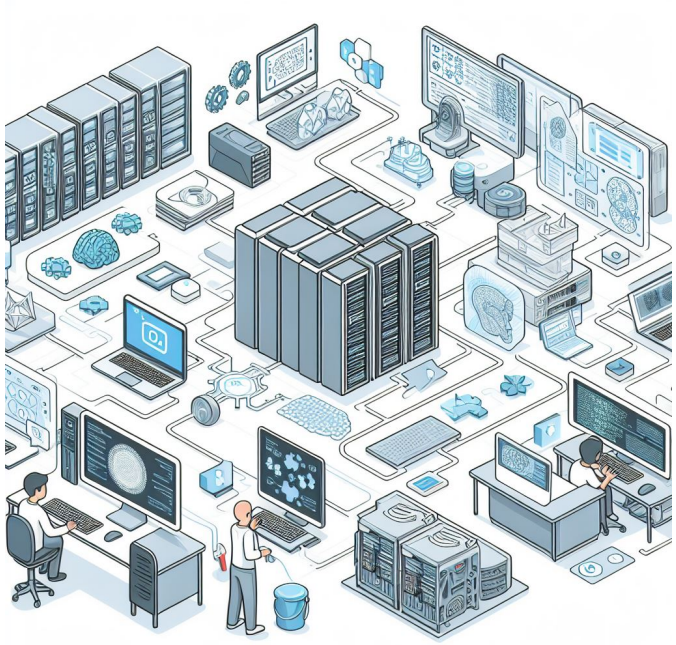
L'inference è il processo di utilizzo di un modello addestrato per fare predizioni o analizzare nuovi dati. Durante l'inference, il modello applica le conoscenze acquisite durante il training o il finetuning per produrre risultati utili senza apportare modifiche ai suoi parametri.

Training, fine tuning e inference in ambienti HPC

Training
Utilizzando l'HPC, è possibile accelerare il processo di addestramento dei modelli AI, riducendo il tempo necessario per raggiungere risultati significativi.

Finetuning
L'HPC permette di eseguire rapidamente iterazioni multiple di finetuning sui modelli esistenti, consentendo un rapido aggiornamento in risposta a nuovi dati o requisiti.

Inference
Con l'HPC, è possibile eseguire inferenze in tempo reale su grandi quantità di dati, garantendo una risposta istantanea alle richieste dell'applicazione senza compromettere le prestazioni.



EC for AI Factory

EC strategy for Large Artificial Intelligence (AI) models

I modelli di AI di grandi dimensioni (LLM) rappresentano una nuova ondata di modelli di Generative AI adattabile a vari domini e compiti.

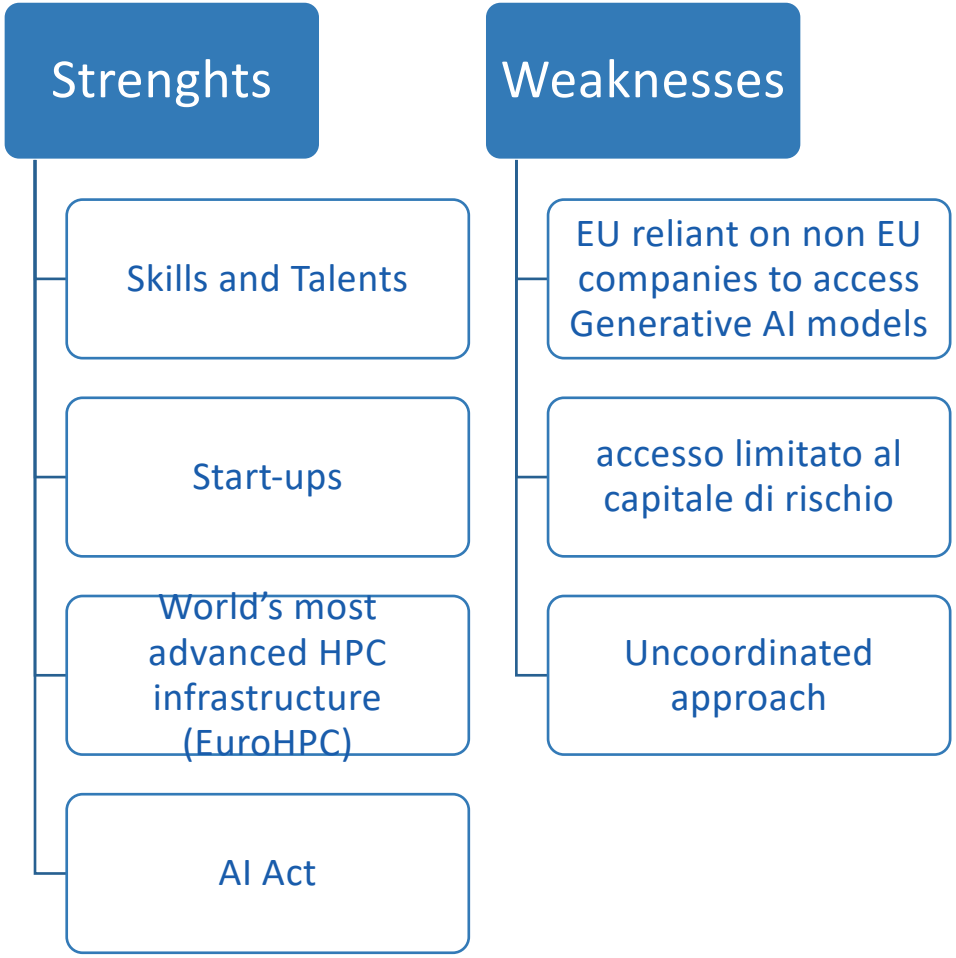
- I modelli di Generative AI hanno un potenziale immenso per rivoluzionare diversi settori.
- La maggior parte dei modelli di IA di grandi dimensioni (ad esempio ChatGPT) non sono europei.
- La padronanza di questa tecnologia è di importanza strategica per l'Europa, in linea con la sicurezza economica.

EC strategy for Large Artificial Intelligence (AI) models

- In her 2023 State of the Union address, President von der Leyen announced that **the supercomputing resources of the EuroHPC JU will be made available to European AI startups to train their large-scale models**, contributing to the EU's aim of leading global advances in AI and of achieving responsible and ethical innovation.

European Commission

Towards large AI models: European strengths and weaknesses



Come sfruttare i punti di forza di EuroHPC per sviluppare un ecosistema di start-up e ricerca sull'AI altamente competitivo?

MA:
- L'attuale infrastruttura e i servizi EuroHPC non sono ottimizzati per l'IA.
- Non si tiene conto delle esigenze specifiche delle comunità di utenti dell'IA.
- Nessuna connessione con l'ecosistema dell'IA.

Amendments to EuroHPC Regulation (EU) 2021/1173

AI Factory

Hosting Entity

- Acquisizione e gestione di supercomputer dedicati all'IA (in co-locazione con il data center).
- Aggiornamento dei sistemi di EuroHPC esistenti con partizioni AI
- Fornire l'accesso alle risorse HPC alle PMI e alle start-up

Centro servizi per AI

- Centro di servizi di supercalcolo per l'IA (algoritmi, formazione, test, valutazione e convalida di modelli di IA, sviluppo di applicazioni di IA su larga scala, ...).
- Strutture per la programmazione di facile utilizzo del supercomputer (parallelizzazione, ottimizzazione dell'uso, ...)

Ecosistema AI

- Attrarre e mettere in comune i talenti
- Interagire con l'ecosistema dell'IA in generale e con altre iniziative di IA.

Come accedere alle risorse computazionali?

MODALITA' DI ACCESSO A LEONARDO

Scientific Merit

- investimento per attività di ricerca scientifica basata sul merito.
- Le proposte (HPC e AI) sono sottoposte a peer-review esterno, per verificarne il merito scientifico, e a valutazione tecnica, da parte di esperti del CINECA, per verificarne l'idoneità a funzionare bene sulle architetture HPC disponibili.
- Esempi:Tutti i rcercatori Italiani e Europei; STRATEGIC ACTIONS: DestinE Earth

Collaborazioni con SME for OPEN SCIENCE

- collaborazione con mondo dei privati in cui l'output è open source, non ha un valore commerciale
- Esempi: Mistral AI, IGenius per Modello Italia

Industrial High Performance Computers

- La CE ma anche noi hanno deciso di sostenere una azione per l'acquisizione di un sistema di supercalcolo che sarà acquisito con fondi privati e messi a disposizione dei privati


HIGH PERFORMANCE COMPUTING
SERVIZI PER LE UNIVERSITÀ E LA RICERCA
SISTEMI INFORMATIVI PER I MINISTERI
IGENIUS E CINECA ANNUNCIANO "MODELLO ITALIA"
Modello Italia ha l'obiettivo di realizzare un nuovo modello GPT, completamente italiano, per aiutare aziende e Pubblica Amministrazione a sfruttare pienamente i vantaggi derivanti dall'Intelligenza Artificiale generativa, anche in settori sensibili come sanità, finanza, sicurezza nazionale. Tutto nel massimo rispetto delle normative sulla privacy e sulla sicurezza nazionale.

Industrial High Performance Computers

The European High Performance Computing Joint Undertaking (EuroHPC JU)

CALL FOR EXPRESSION OF INTEREST for the selection of consortia of private partners and the Hosting Entities for the procurement of Industrial High Performance Computers

CALL FOR PROPOSALS

CALL FOR EXPRESSION OF INTEREST for the selection of consortia of private partners and the Hosting Entities for the procurement of Industrial High Performance Computers

The overall objective of this call is to select hosting entities for industrial-grade supercomputers which will be acquired by the EuroHPC JU together with a consortium of private partners.


Grazie
Gabriella Scipione
g.scipione@cineca.it
CINECA