Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

terabit

TeRABIT, un'autostrada di elaborazione dati per la ricerca

Stefano Salon

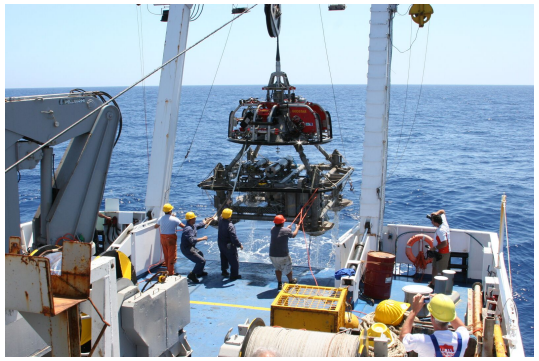
29 Settembre 2024

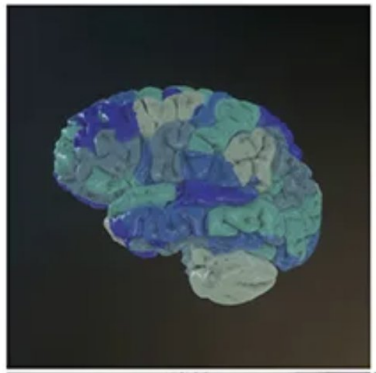
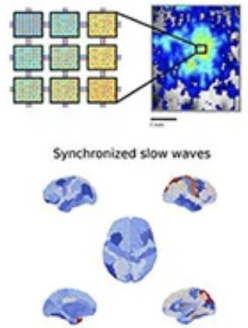
Trieste NEXT

Sommario

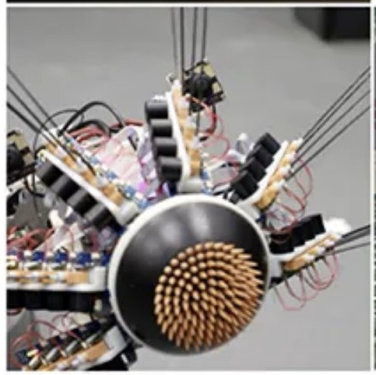
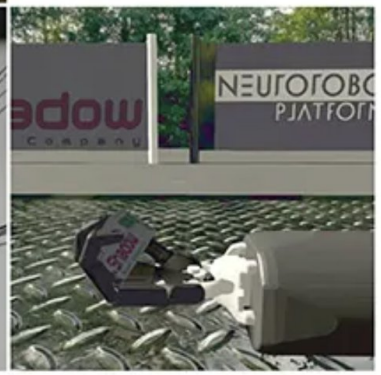
- Cosa sono le Infrastrutture di Ricerca
- L'importanza dei dati nella ricerca (e non solo) e il ruolo delle Infrastrutture di Ricerca
- Il Progetto TeRABIT in poche parole nel contesto del PNRR
- La sinergia fra 3 Infrastrutture di Ricerca: 1+1+1 = 3 ?
- Gli impatti per i ricercatori e le opportunità per le aziende private

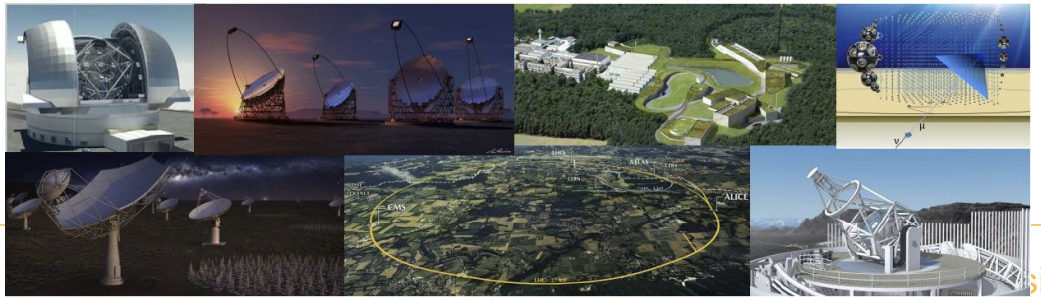
Le Infrastrutture di Ricerca (IR) Europee



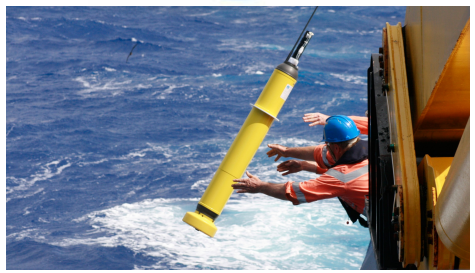





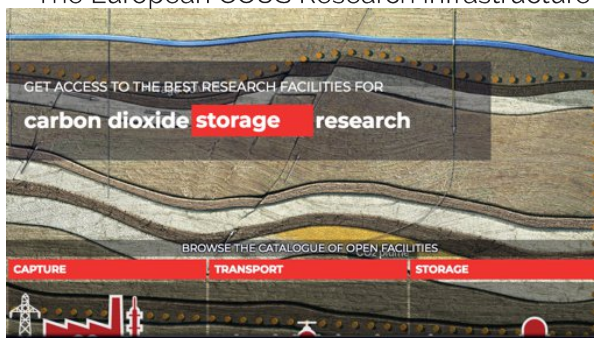



Le Infrastrutture di Ricerca (IR) EU coordinate da OGS



International program that uses profiling floats to observe oceans
www.euro-argo.eu

The European CCUS Research Infrastructure

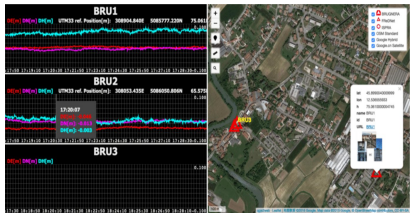


European Carbon Dioxide Capture and Storage Laboratory Infrastructure
www.eccsel.org

Partnership for Advanced Computing in Europe
www.prace-ri.eu

European Strategy Forum on Research Infrastructures

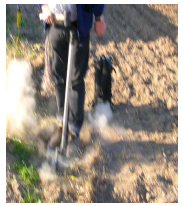
Altre Infrastrutture di Ricerca, osservatori e laboratori di OGS

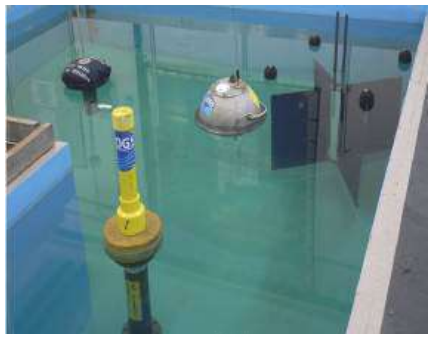
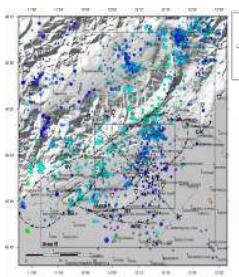
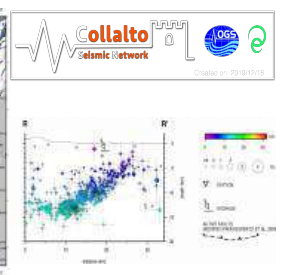

















BIG
DATA
STORAGE
VOLUME
PROCESSING
DECISION
INFORMATION
VARIETY
VELOCITY
SPEED
OPTIMIZATION
SHARING
COLLECTION
ANALYSIS
TOOLS
TRANSFER
CAPTURE
DIFFICULT
DATABASE
PROCESS
SEARCH
APPLICATIONS
EXPLORE
CHALLENGES
SIMULATION
RELATIONAL
SOURCES
SYSTEMS
VISUALIZATION
SOFTWARE
BUSINESS
STATISTICS
PETABYTES
AMOUNT
METEOROLOGY
CURATE
DOMAIN
MANAGE
GOVERNMENT
RANGE
COMPLEX
ASSETS
DEPENDENCIES
STREAM
ENABLE
TRIGGER
ENHANCED
CAPABILITIES
LOGS
TYPES
HIGH
TRENDS
INSIGHT
DENSITY
MAKING
NETWORKS
SCIENTISTS
RESEARCH
SIMULATIONS
DISCOVERY
INTERNET
TERABYTES
SIZE


BIG DATA
NETWORK
COLLECTION
STORAGE
RESEARCH
ANALYSIS
VOLUME
EXABITE
VISUALIZATION
CLOUD TECHNOLOGY

Un diluvio di dati… ma come li misuriamo?

Un disco esterno da 1 Terabyte (TB, 10^{12} byte) costa 50-100 EUR e consente di archiviare circa:

- **250.000 foto scattate con una fotocamera da 12 MP, OPPURE**
- **250 film o 500 ore di video in HD, OPPURE**
- **6,5 milioni di pagine di documenti (es. file Office, PDF, presentazioni…)**

1 TB di spazio corrisponde all'incirca a:

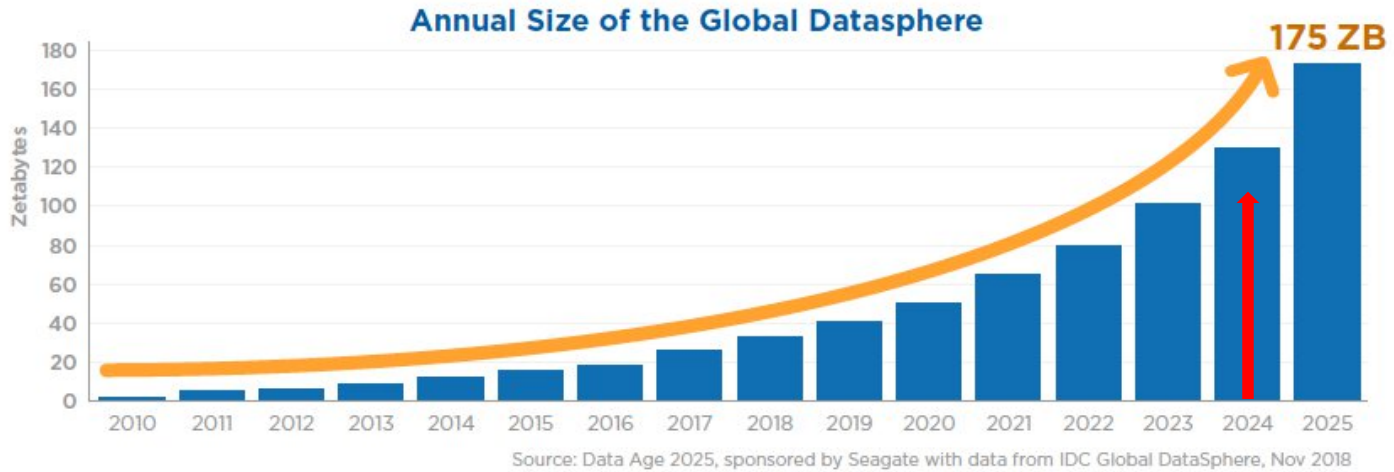
- **16 iPhone o Samsung Galaxy da 64 GB**
- **4 laptop Windows o MacBook da 256 GB**
- **1.300 archivi fisici di file cartacei**

Fonte: Dropbox

1 Zettabyte (ZB, 10^{21} byte) = 1 miliardo di Terabyte … mettendoli in fila copriamo quasi ¼ della distanza Terra-Luna

Un diluvio di dati: la *materia prima* dell'era dell'informazione

Figure 1 – Annual Size of the Global Datasphere



2010

Seagate-IDC (2018), 2012: 1 Zettabyte of global datasphere (2.5 Exabyte of marketing data)

Today, Large Language Models (LLMs) datasets are constantly growing, e.g. the Common Crawl dataset (250 billion pages since 2017, 3-5 new pages each month) grows monthly of 4-500 TB and has reached O(1000 TB)

Un diluvio di dati
Il ruolo delle Infrastrutture di Ricerca

Le IR producono moltissimi **dati per la ricerca**, p.es. monitoraggi ambientali, proiezioni climatiche, esperimenti di laboratorio, dati astronomici, dati degli acceleratori…, e alcune IR *sono* banche dati

Ogni disciplina sia scientifica che umanistica ha il **DATO** come valore fondamentale (Data Centric)

Questo richiede:

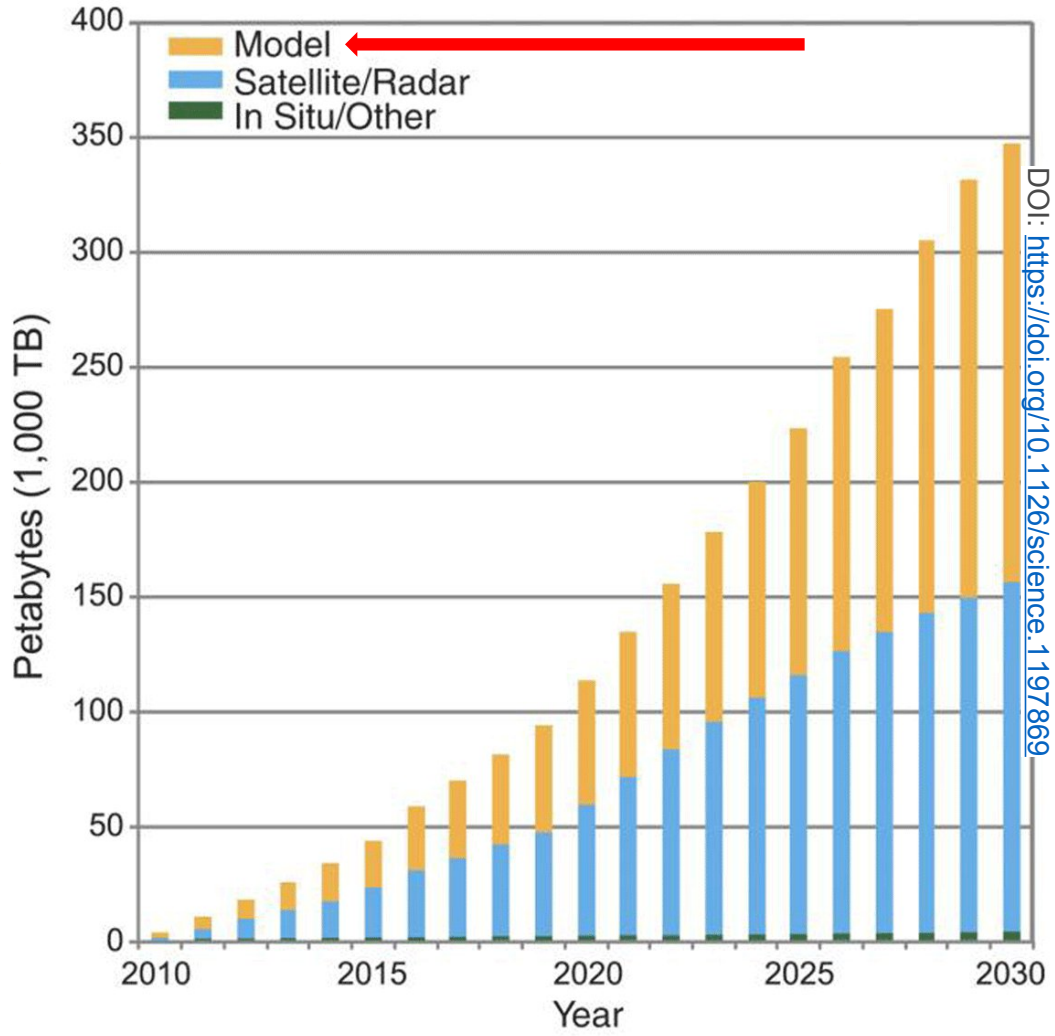
- Gestione e analisi dei dati che necessitano **grandi risorse IT** con software e hardware in rapida evoluzione
- Sistemi di calcolo ad alte prestazioni per **integrare i dati con i modelli** e produrre previsioni o scenari
- Reti di comunicazione ultraveloci fino al **Terabit al secondo** (1 Tbps = 1.000 miliardi di bit al secondo) per traferire i dati istantaneamente dove servono

Un diluvio di dati
Il ruolo delle Infrastrutture di Ricerca

La Earth System Grid Federation (**ESGF**) che raccoglie i dati degli scenari climatici usati dal IPCC per redigere i rapporti ha avuto un aumento enorme del volume di dati climatici dal 2007 (CMIP3: 40 TB, CMIP5: 2000 TB, CMIP6: 20000 TB)

IPCC: Gruppo Intergovernativo per il Cambiamento Climatico

CMIP: Progetto di comparazione fra modelli climatici



TeRABIT nel contesto PNRR

MUR - Direzione generale dell'internazionalizzazione e della comunicazione

Avviso pubblico n.3264 del 18.12.2021

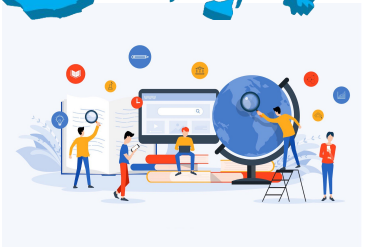
Missione 4, "**Educazione e Ricerca**"

Componente 2, "**Dalla ricerca all'impresa**"

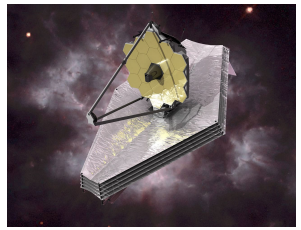
Linea di investmento 3.1, "Fondo per la realizzazione di un **sistema integrato di infrastrutture di ricerca e innovazione**"

Azione 3.1.1 Creazione di nuove IR o potenziamento di quelle esistenti che concorrono agli obiettivi di Eccellenza Scientifica di Horizon Europe e costituzione di reti






apply research to production



RIs

TeRABIT: i partner del progetto e il finanziamento

INFN Istituto Nazionale di Fisica Nucleare

OGS Istituto Nazionale di Oceanografia e di Geofisica Sperimentale

Consortium GARR

CINECA

41 milioni EUR, di cui:

- ✓ 35 per le infrastrutture
- ✓ 4 per il personale

Sfruttare la complementarietà di tre IR già in funzione riconosciute come prioritarie dal PNIR (2021-2027)

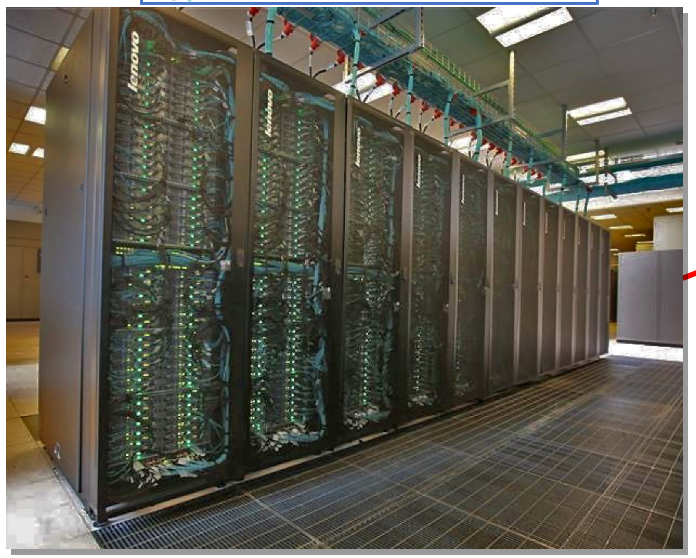
stato delle IR nel 2023

EuroHPC Joint Undertaking

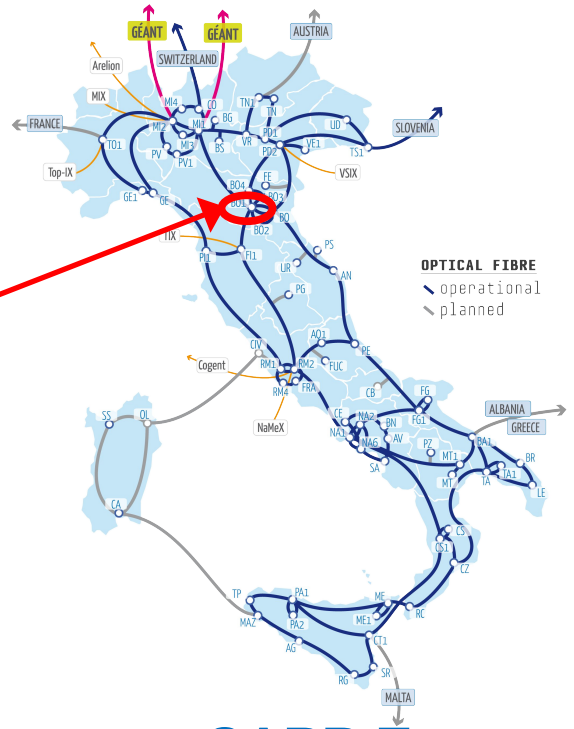
PRACE PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

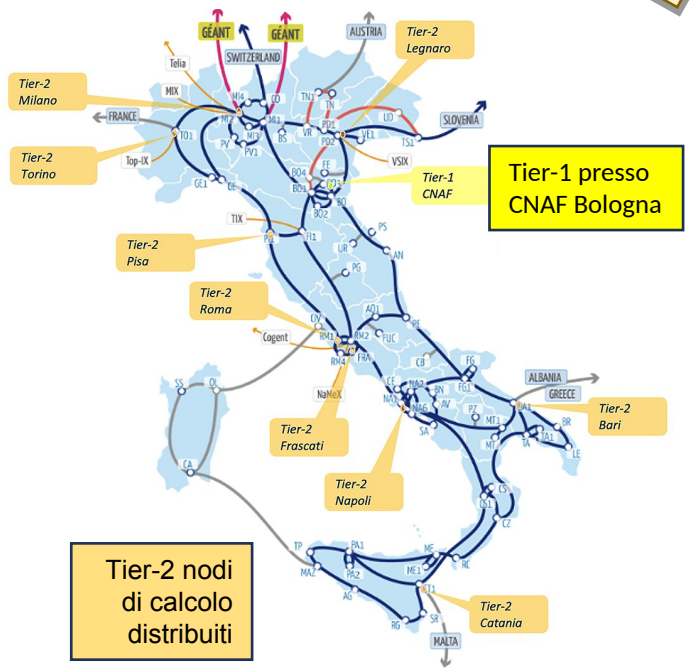
GÉANT

WLCG Worldwide LHC Computing Grid



Galileo100 - HPC presso CINECA (Bologna)





PRACE-Italy **GARR-T** **HPC-BD-AI**

Visione

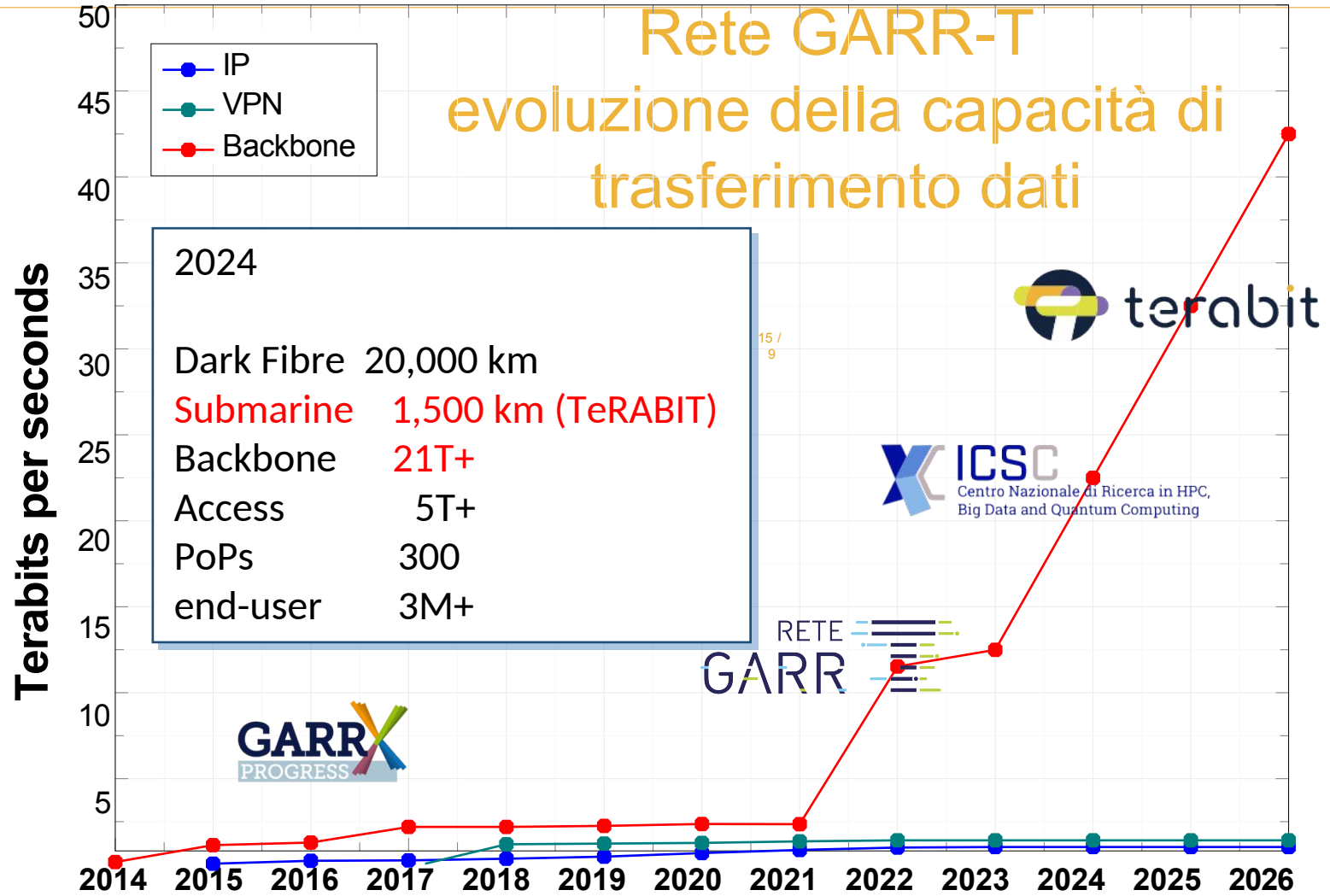
Creare un ambiente digitale HPC-Cloud ibrido, distribuito e iperconnesso che:

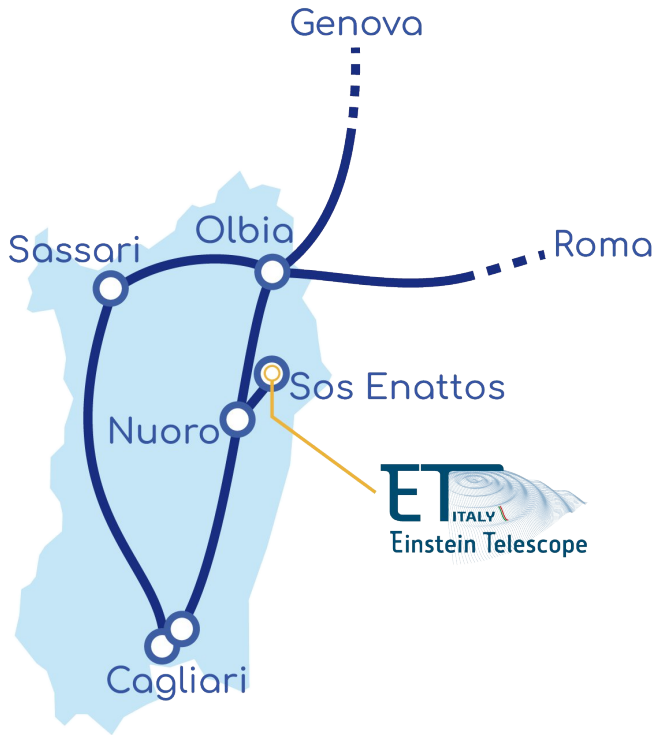
- offra servizi progettati per soddisfare le esigenze in continua evoluzione di ricerca e innovazione
- federi e rafforzi le tre IR esistenti GARR-T, PRACE-Italy e HPC-BD-AI (HPC-Big Data-Artificial Intelligence) in sinergia con il Centro Nazionale per il Supercalcolo, i Big Data e il Quantum Computing - ICSC
- sfrutti le loro connessioni esistenti con altre IR e data space nazionali ed europei attraverso il backbone della rete Europea ad alta capacità GÉANT

Obiettivi principali

1. Abilitare un trasferimento di dati diffuso, fino ai Terabits per secondo, e servizi su scala nazionale in Italia, con particolare attenzione alle regioni meridionali e insulari, tutte connesse alla rete Europea
2. Innovare il nodo centrale HPC di PRACE-Italy, mantenendo lo stesso livello di prestazioni (Tier-1)
3. Innovare i servizi HPC offerti ai ricercatori, oltre il modello di calcolo centralizzato, aggiungendo

Rete GARR-T
evoluzione della capacità di trasferimento dati



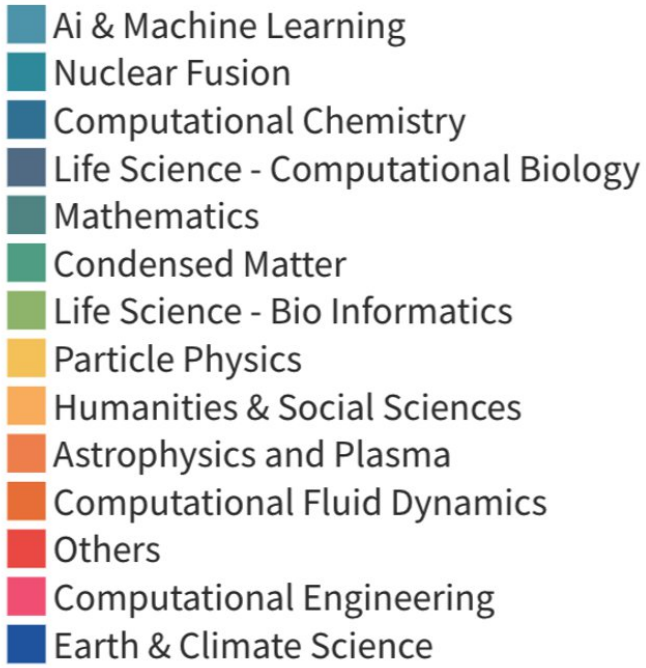


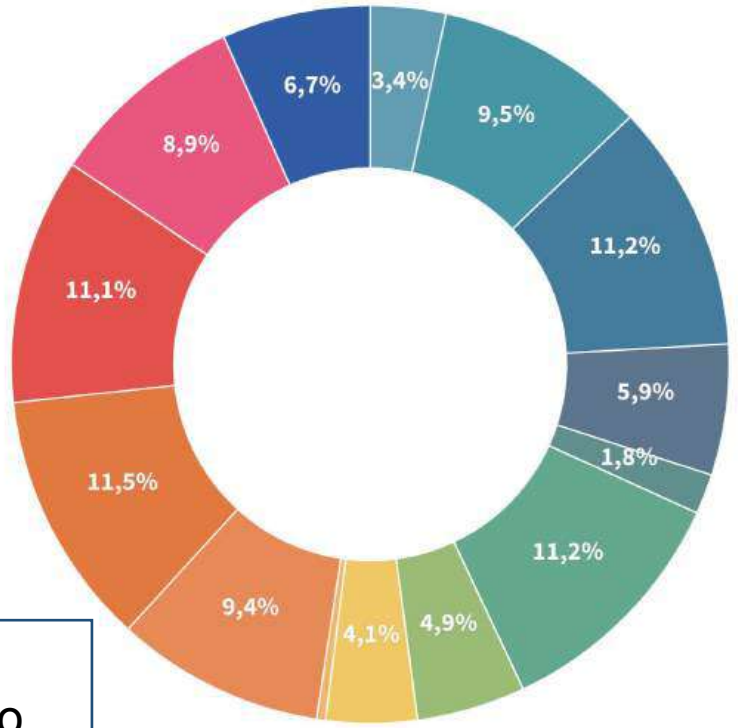
terabit

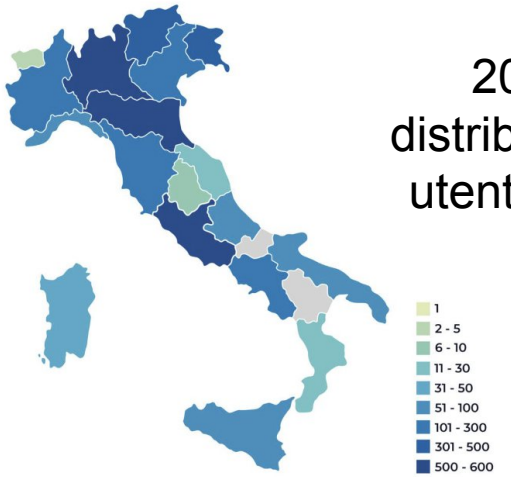
Scientific domains

Scientists use Cineca computational resources within all scientific disciplines. The most represented three are Computational Chemistry, Condensed Matter Physics and Computational Fluid Dynamics, with about11% each, followed by Nuclear Fusion (10%), Computational Engineering, Astrophysics, and Plasma Physics with more than 9% each.

Uso e potenziamento di PRACE-Italy







2022 distribuzione utenti HPC

TeRABIT upgrade	AUMENTO
Capacità HPC (n. nodi)	~ x 5.0
Storage	~ x 2.5

Nel 2022, dopo una rigorosa revisione, le risorse allocate sono state del 110%

…sulla base delle richieste degli utenti

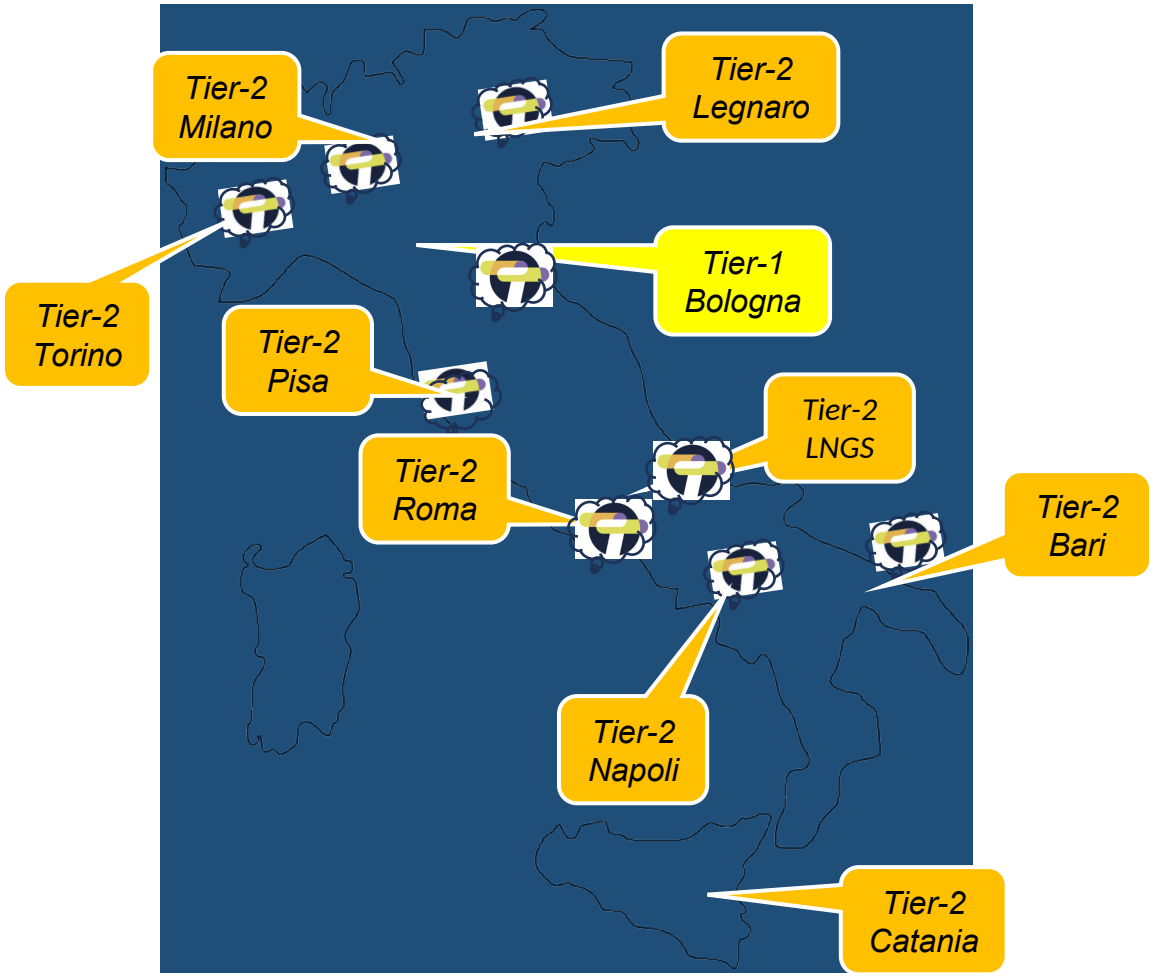
Evoluzione di HPC-BD-AI

Le HPC-Bubbles si aggiungono alla IR cloud distribuita di INFN, come nodi di calcolo molto compatti ma potenti e caratterizzati da diverse tipologie HW

Nodi: Tipo 1 : CPU (192 cores)
Tipo 2 : CPU + GPU (4x NVIDIA H100)
Tipo 3 : CPU + FPGA

Siti: CNAF, Bari, LNGS, Milano Bicocca, Napoli, Padova, Pisa, Roma 1, Torino

Storage aggiuntivo:
Mass storage : CNAF
Storage ad alte prestazioni : CNAF, Bari



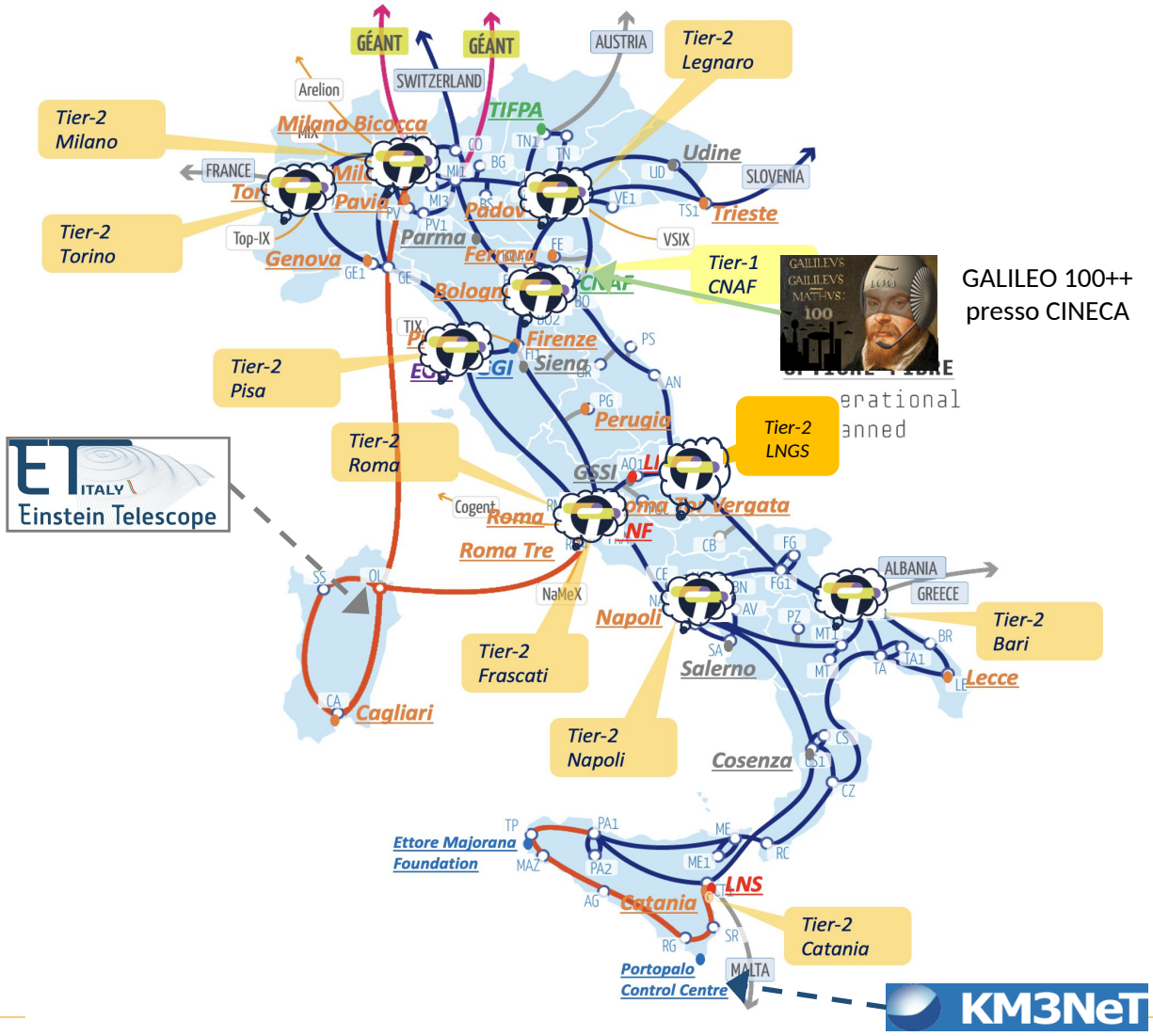
Le IR a fine progetto (2025)

L'immagine mostra la sovrapposizione prevista delle topologie fisiche combinate di tutte le tre IR a fine progetto:

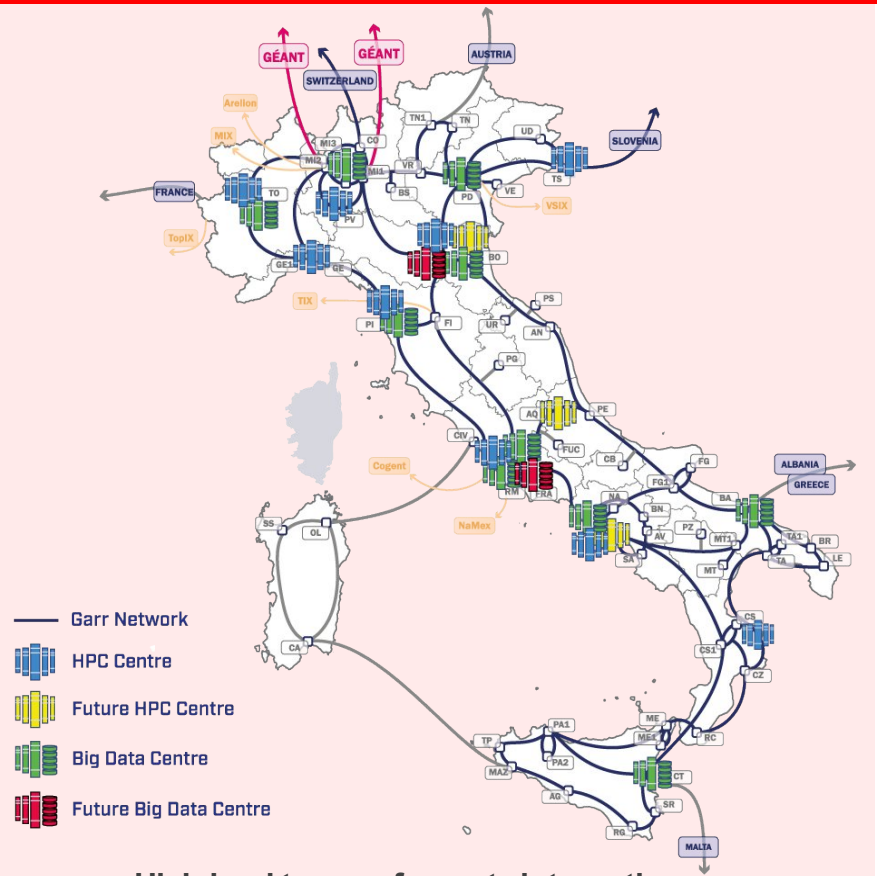
- GARR-T con (in rosso) le nuove fibre (isole)
- HPC-BD-AI con i siti delle HPC-Bubbles
- PRACE-Italy con il potenziamento di GALILEO100 presso il CINECA

Gli sviluppi sono in sinergia con ICSC

L'estensione della rete di TeRABIT darà supporto ad altre IR, p.es. ET (onde gravitazionali) e KM3NeT (neutrini)



0 SUPERCOMPUTING INFRASTRUCTURE



High-level teams of experts integrating the Spokes working groups (mixed cross-sectional teams)

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

ICSC è composto da

- 10 sotto-progetti (spoke)
- 1 spoke per la parte infrastrutturale

ISTRUZIONE E FORMAZIONE, IMPRENDITORIALITÀ, TRASFERIMENTO DI CONOSCENZE, POLICY, OUTREACH

1 FUTURE HPC & BIG DATA

2 FUNDAMENTAL RESEARCH & SPACE ECONOMY

3 ASTROPHYSICS & COSMOS OBSERVATIONS

4 EARTH & CLIMATE

5 ENVIRONMENT & NATURAL DISASTERS

6 MULTISCALE MODELING & ENGINEERING APPLICATIONS

7 MATERIALS & MOLECULAR SCIENCES

8 IN-SILICO MEDICINE & OMICS DATA

9 DIGITAL SOCIETY & SMART CITIES

10 QUANTUM COMPUTING

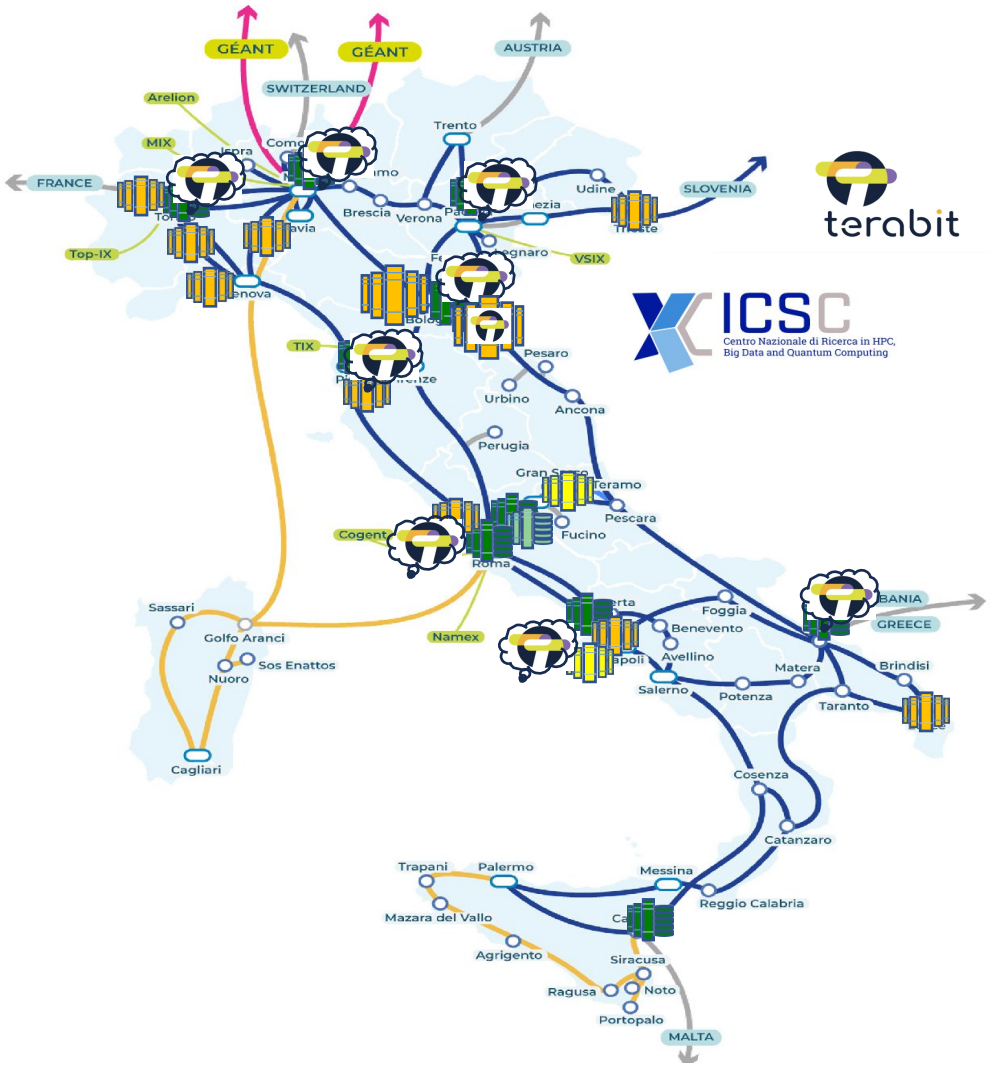
Una IR nazionale per il supercalcolo

ICSC e TeRABIT mirano a creare una IR nazionale federata per il supercalcolo

L'accesso alle risorse sarà trasparente per l'utente finale

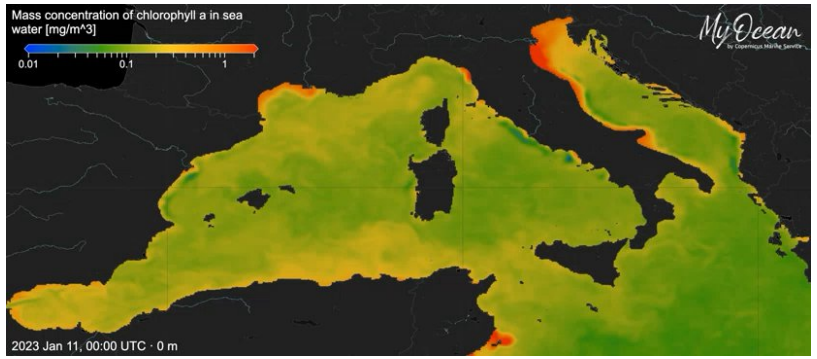
Gli attori principali sono: INFN, CINECA, GARR, OGS

E inoltre: CMCC, ENEA, SISSA, IIT, Univ. TO, Univ. Roma Sapienza, Univ. TS, INAF, CNR, …

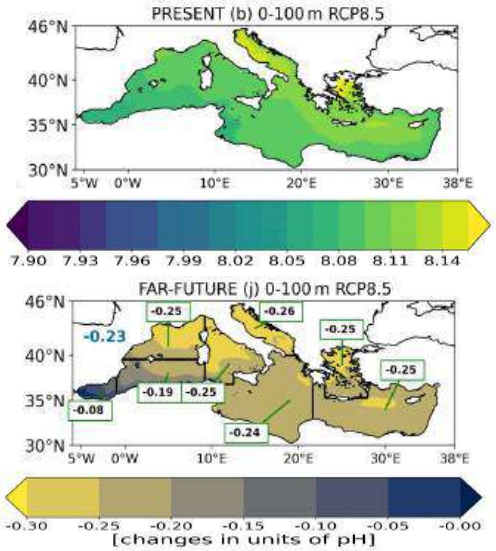


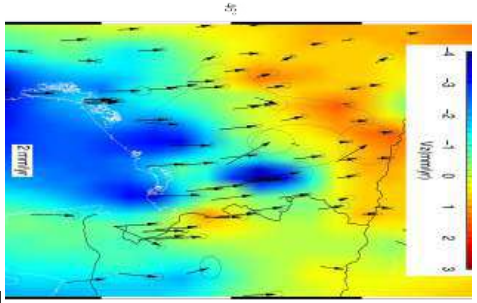
OGS utente TeRABIT

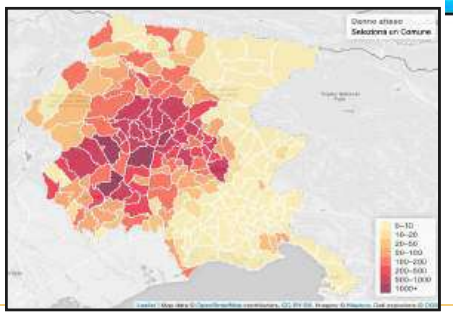
- Oceanografia operativa e digital twin nel Mar Mediterraneo e nel Nord Adriatico nel contesto del servizio EU Copernicus
- Scenari di cambiamento climatico ed effetti multi-scala sugli ecosistemi marini, costieri e lagunari
- Modellazione del sistema terrestre per l'analisi del ciclo del carbonio
- Monitoraggio sismico regionale (anche tramite elaborazione dati GNSS), valutazione probabilistica del rischio e produzione di scenari di danno
- Simulazione 3D della propagazione delle onde sismiche in strutture geologiche complesse

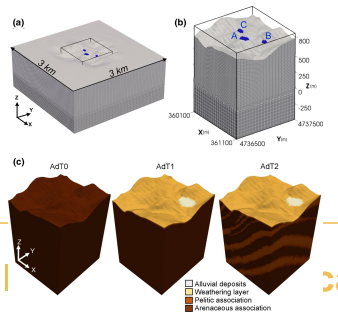


OGS



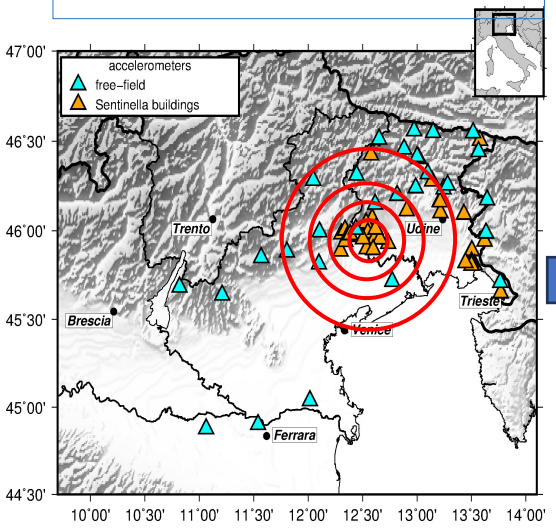






Caso studio: SHAKE (Supercomputing for eartHquAKE rapid damage assessment)

Rete OGS SMINO
Sensori al suolo e sugli edifici (oggi oltre 400+)

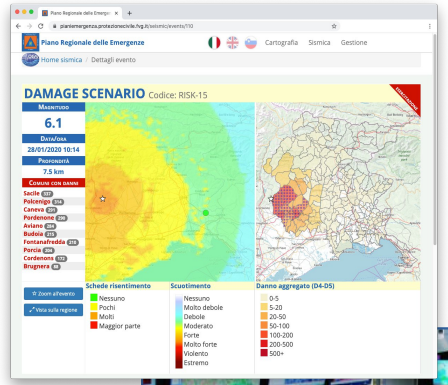


HPC

Codice physics-based per la simulazione del moto del terreno con informazioni di dettaglio per fornire

Valutazione delle scosse sismiche

Valutazione dell'impatto



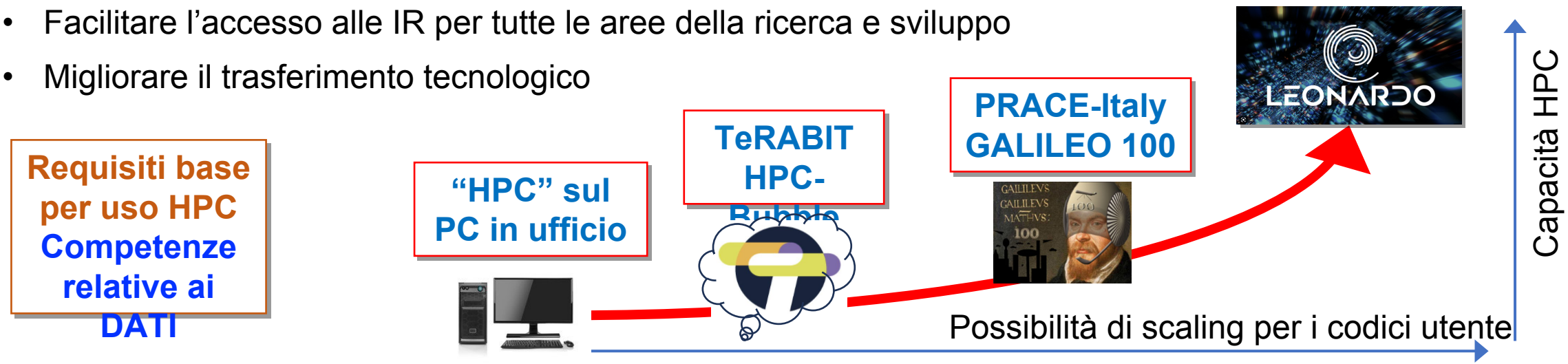
Protezione Civile

Challenge :

Usare le infrastrutture HPC per fornire una valutazione rapida del danno

Obiettivi - Innovazione - Impatti

- Potenziamento delle IR per rispondere alle esigenze di scalabilità e alle nuove esigenze della ricerca
- Maggiore integrazione tra servizi di rete, dati e HPC con servizi comuni e federati
- Servizi HPC innovativi ("Bubbles"), e capacità HPC/AI modulare e crescente tra l'"edge", dove si trovano gli utenti e i loro dati, e PRACE-Italy, in sinergia con ICSC (Leonardo)
- Federazione e comunicazione fra infrastrutture HPC in stretta collaborazione con altri centri nazionali e internazionali (via GÉANT) come PRACE e i siti di EuroHPC
- Facilitare l'accesso alle IR per tutte le aree della ricerca e sviluppo
- Migliorare il trasferimento tecnologico



Un progetto che cresce nelle comunità della ricerca

Master HPC e PhD

Studenti

Education

Young technologists and researchers

Other

Case study: Earth Science

Analisi dei terremoti

Sviluppi in stretta collaborazione con

ICSC National Centre for High Performance Computing, Big Data and Quantum Computing

Utenti: educazione e disseminazione

Casi studio

Potenziamento della formazione e della ricerca HPC per le Scienze della Terra e partecipazione di esperti di altre aree di ricerca

formazione di giovani tecnologi

- organizzazione di **2 workshop (1st in June 2**
- organizzazione di **1 hackathon**

Master HPC e PhD

Studenti di dottorato (**11**) e di master HPC (**7**)

Studenti

- Sessioni per presentare il progetto TeRABIT project nelle scuole superiori distribuite in Italia
  - ✓ Formare i ricercatori e i tecnologi del futuro

Grazie. Domande ?

For information: info@terabit-project.it

Trends in research (Big) Data: the role of Research Infrastructures

DATA LAKE LEVEL

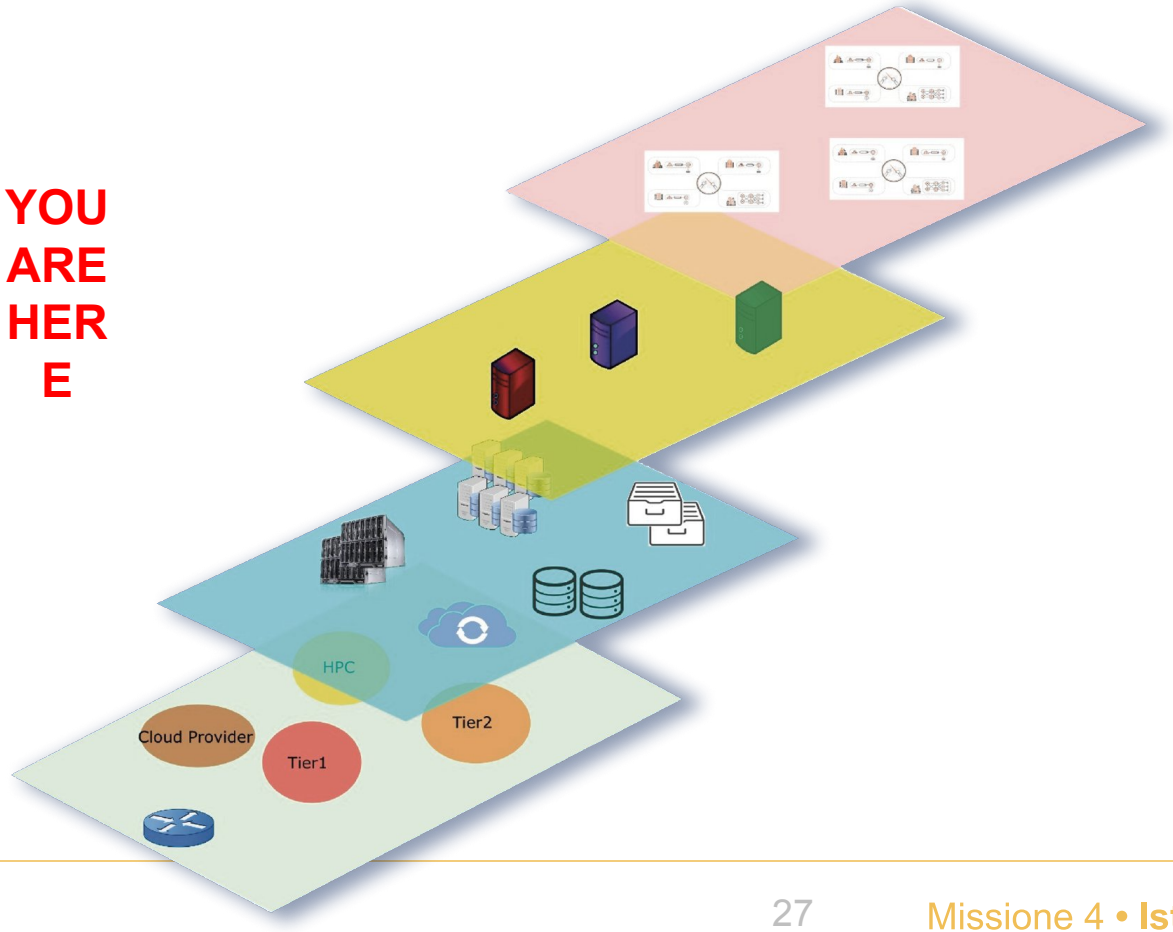
USER COMMUNITY LEVEL

YOU ARE HERE

GENERAL SERVICES LEVEL

INFRASTRUCTURE LEVEL

A Data Lake for research: the high-level view



Data Lake entry point & customized data services

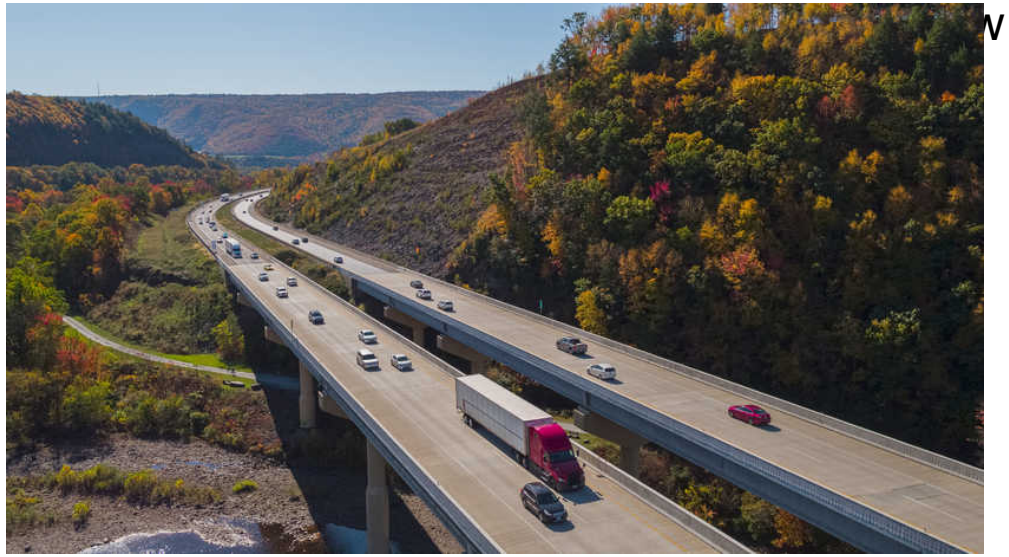
Individual user communities services

Generic services

Data centres and resource providers

Planned impact of the project: the user view

- NOW: user logs in on G100 => AFTER: user will log in on the “federated layer” learning new tools and exploiting new services, hiding the complexity of the infrastructure behind
- TeRABIT is building up the new services and the knowledge support (workshops, hackathons, training events…)
- Services are the project legacy: they are built on Research Infrastructures and will remain after TeRABIT
- w paradigm…



The highway is GARR-T

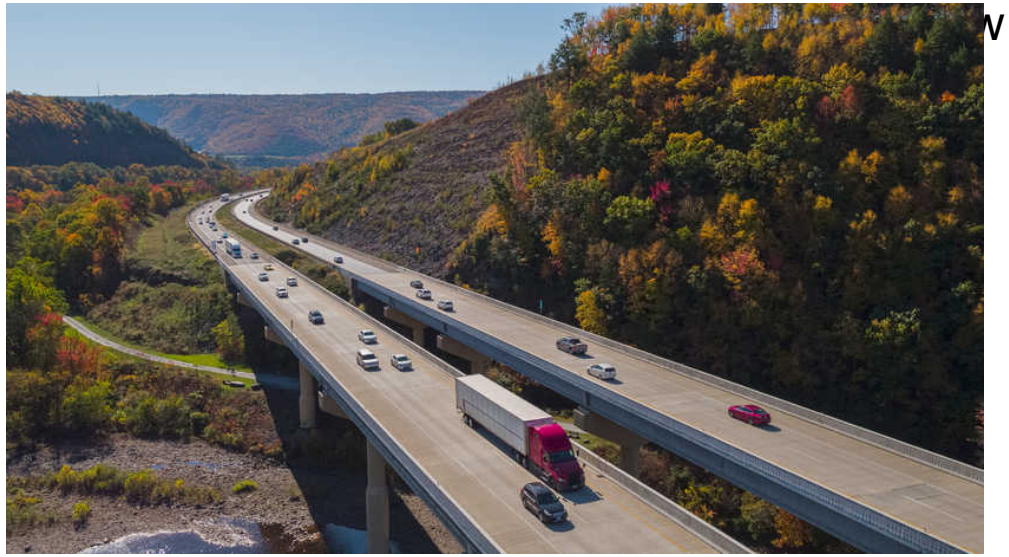
VANS are VMs on Cloud

BIG TRUCKS are INFN Bubbles

FERRARI is Cineca HPC

Planned impact of the project: the user view

- NOW: user logs in on G100 => AFTER: user will log in on the "fe
  exploiting new services, hiding the complexity of the infrastructur
- TeRABIT is building up the new services and the knowledge supp
  events…)
- Services are the project legacy: they are built on Research Infras
- w paradigm…



The highway is GAR

VANS are VMs on C

BIG TRUCKS are INF

FERRARI is Cineca H