Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

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

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

INFN
Istituto Nazionale di Fisica Nucleare

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

Introduction to High Rate analysis

Tommaso Diotalevi (UniBO)

Workshop on "Quasi-Interactive Analysis of Big Data with High Throughput" - 8/10 Jan 2024 - Bologna

ICSC: The National Center for HPC, Big Data and Quantum Computing

what is it?

- Italy has funded, with NRRP (pandemic recovery) funds, **5 large National Centers**, for a total of **1.6B €** over 3 years, on key future technologies.
- One of them, coordinated by **INFN**, focuses on modern IT technologies, with the final goal of deploying a <u>long-term distributed infrastructure</u> (>> 3y) <u>for national research and industrial development</u>.
- The project started on September 2022, lasting until December 2025.

1. **ICSC: Big Data, High Performance and Quantum Computing** — ICSC
2. **Advanced Technologies in agricolture (Agritech)** — agritech, National Research Center for Technology in Agriculture
3. **Sustainable Mobility** — MOST
4. **Development of RNA drugs and gene therapies** — RNA &GENETHERAPY
5. **Biodiversity** — NATIONAL BIODIVERSITY FUTURE CENTER

The participants

The ICSC foundation: public and private members

an initiative spread across Italy

National institutes

INFN Istituto Nazionale di Fisica Nucleare

CINECA

ENEA

Consiglio Nazionale delle Ricerche

INAF ISTITUTO NAZIONALE DI ASTROFISICA

ISTITUTO NAZIONALE DI GEOFISICA E VULCANOLOGIA

Consortium GARR

12 Research institutes

HUB

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

UNIVERSITÀ DI PARMA

OGS

Privates

enel

ENGINEERING THE DIGITAL TRANSFORMATION COMPANY

eni

FERROVIE DELLO STATO ITALIANE

FINCANTIERI

fondazione innovazione urbana

autostrade per l'italia

HUMANITAS RESEARCH HOSPITAL

iFAB INTERNATIONAL FOUNDATION BIG DATA AND ARTIFICIAL INTELLIGENCE FOR HUMAN DEVELOPMENT

INTESA SANPAOLO

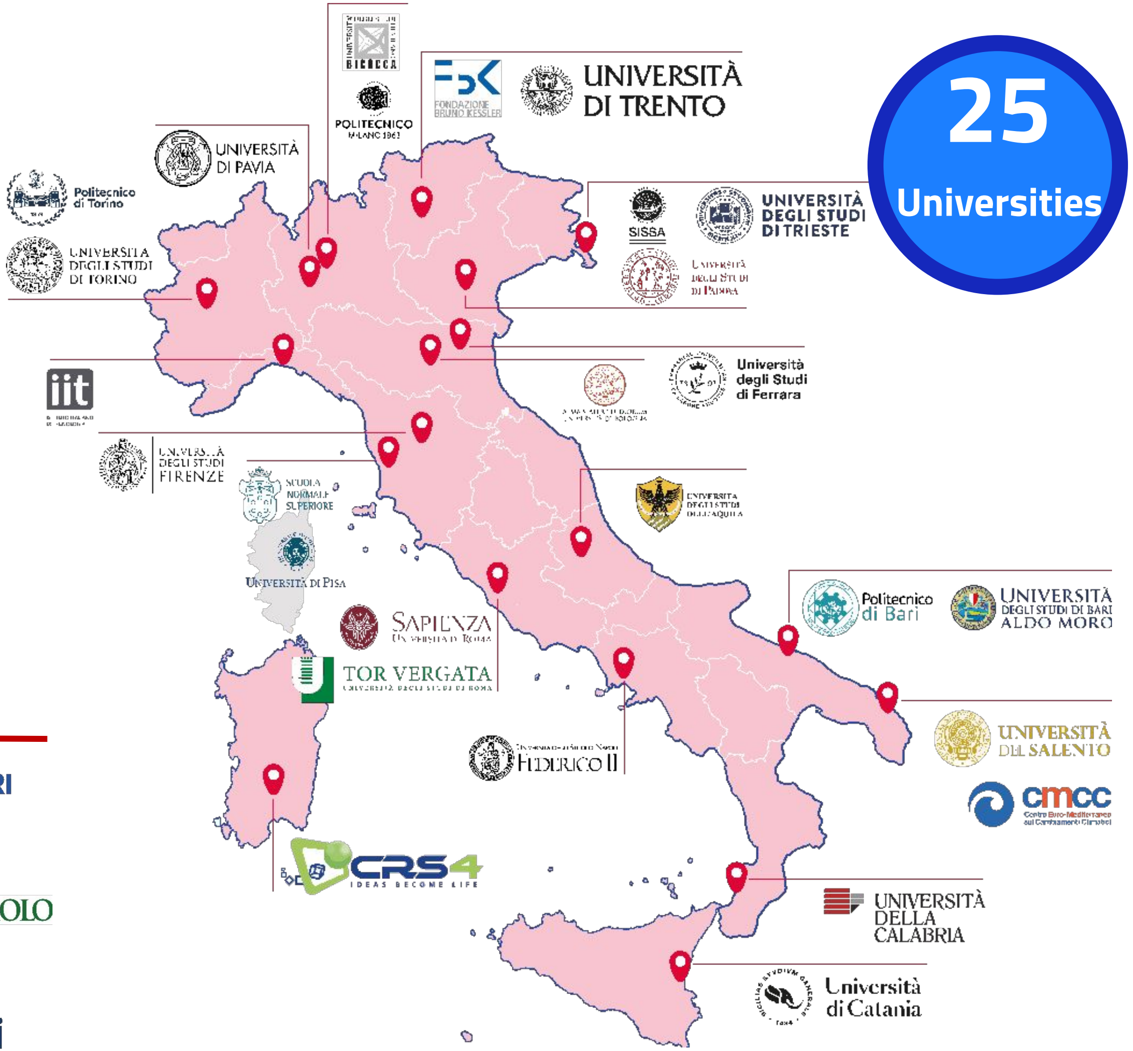
LEONARDO

sogei

ThalesAlenia Space

Terna Driving Energy

UnipolSai ASSICURAZIONI

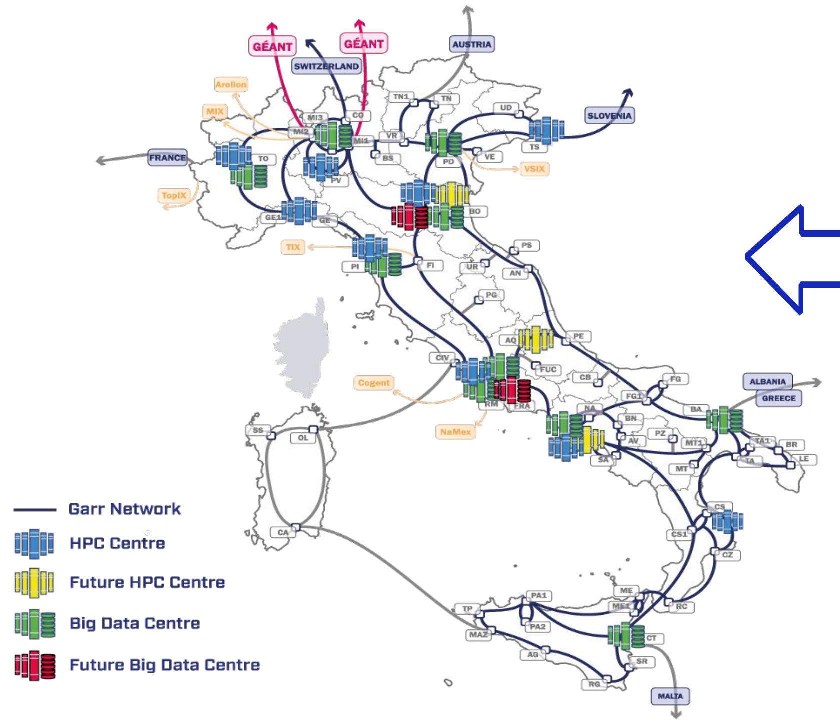

UNIVERSITÀ BICOCCA
POLITECNICO MILANO 1863
FONDAZIONE BRUNO KESSLER
UNIVERSITÀ DI TRENTO
UNIVERSITÀ DI PAVIA
Politecnico di Torino
UNIVERSITÀ DEGLI STUDI DI TORINO
SISSA
UNIVERSITÀ DEGLI STUDI DI TRIESTE
UNIVERSITÀ DEGLI STUDI DI PADOVA
iit
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
Università degli Studi di Ferrara
UNIVERSITÀ DEGLI STUDI FIRENZE
SCUOLA NORMALE SUPERIORE
UNIVERSITÀ DI PISA
UNIVERSITÀ DEGLI STUDI DELL'AQUILA
SAPIENZA UNIVERSITÀ DI ROMA
TOR VERGATA UNIVERSITÀ DEGLI STUDI DI ROMA
Politecnico di Bari
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
UNIVERSITÀ DEL SALENTO
FEDERICO II
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici
CRS4 IDEAS BECOME LIFE
UNIVERSITÀ DELLA CALABRIA
Università di Catania

25 Universities

The structure of the ICSC National Center

The ICSC includes:
10 thematic Spokes and **1 Infrastructure Spoke**

0 SUPERCOMPUTING CLOUD INFRASTRUCTURE



EDUCATION AND TRAINING, ENTREPRENEURSHIP, KNOWLEDGE TRANSFER, POLICY MAKING, 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 APPLICATI

7 MATERIALS & MOLECULAR SCIENCES

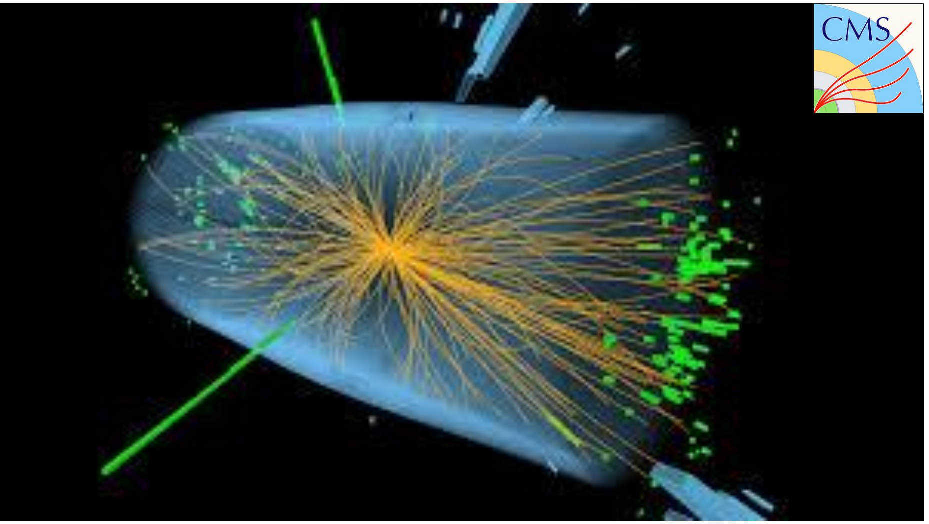
8 IN-SILICO MEDICINE & OMICS DATA

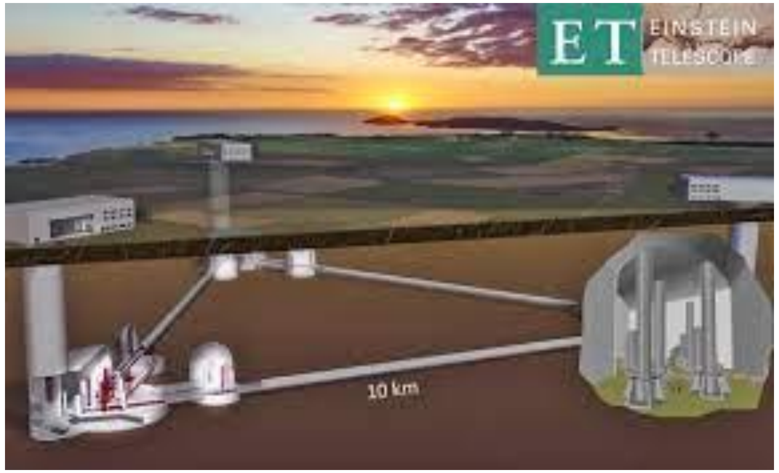
9 DIGITAL SOCIETY & SMART CITIES

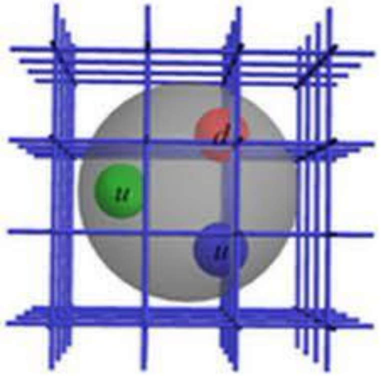
1
0QUANTUM COMPUTING

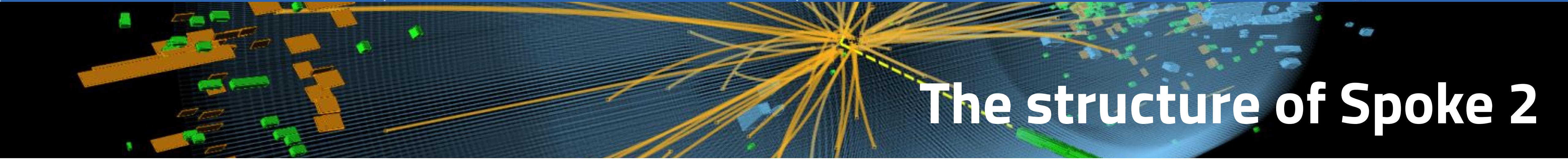
- One Spoke dedicated to building the infrastructure
- Ten thematic Spokes, one of which dedicated to the HEP and Astroparticle research domains.

Spoke 2 – Who are «we»?





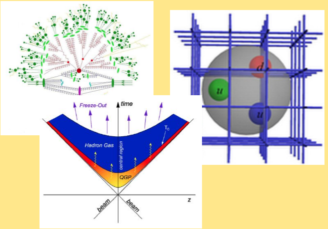




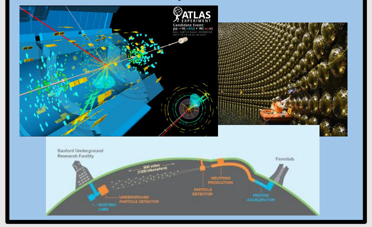
The structure of Spoke 2

Scientific

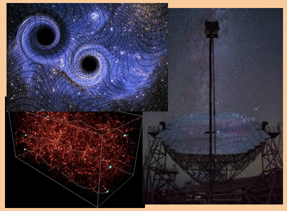
WP1: tools and algorithms for Theoretical Physics



WP2: tools and algorithms for Experimental High Energy Physics



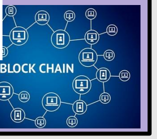
WP3: tools and algorithms for Experimental Astroparticle Physics and Gravitational waves



ICSC-SPOKE2
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

WP6: cross domain initiatives + space economy

GEANT4 A SIMULATION TOOLKIT
copernicus Europe's eyes on Earth



WP5: Boosting computational performance on the distributed CN infrastructure

SCIENTIFIC DATA MANAGEMENT RUCIO
INFN CLOUD
jupyter

WP4: tools for porting/optimization on new architectures (low power, GPU, FPGA, …)

alpaka

NVIDIA. CUDA.

Technologic

FUNDAMENTAL RESEARCH & SPACE ECONOMY

Istitution leader

INFN Istituto Nazionale di Fisica Nucleare

Istitution co-leader

INAF ISTITUTO NAZIONALE DI ASTROFISICA

Istitutions and Universities

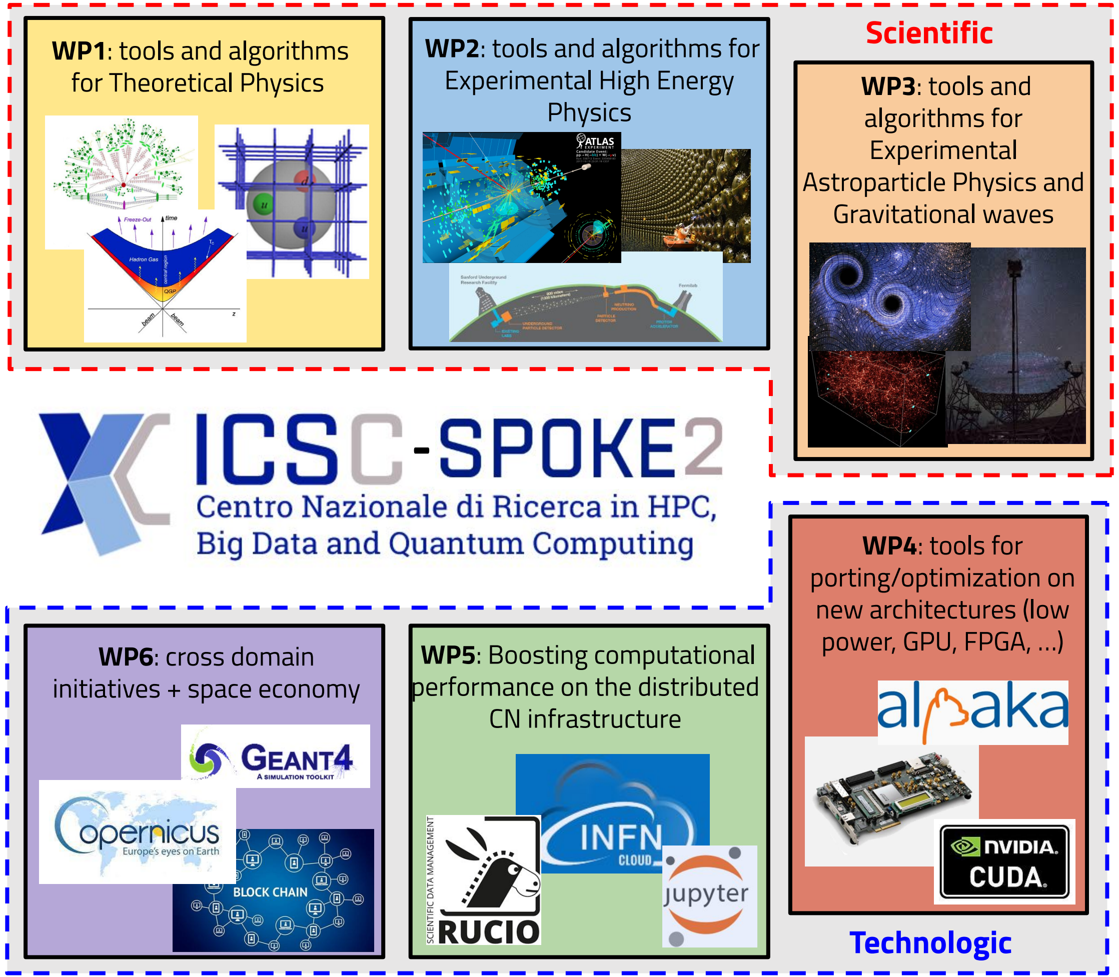
Politecnico di Bari
SAPIENZA UNIVERSITÀ DI ROMA
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
UNIVERSITÀ DELLA CALABRIA
Università di Catania
Università degli Studi di Ferrara
UNIVERSITÀ DEGLI STUDI FIRENZE
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II
UNIVERSITÀ DEGLI STUDI DI PADOVA
UNIVERSITÀ DEL SALENTO
UNIVERSITÀ DEGLI STUDI DI TRIESTE

Companies

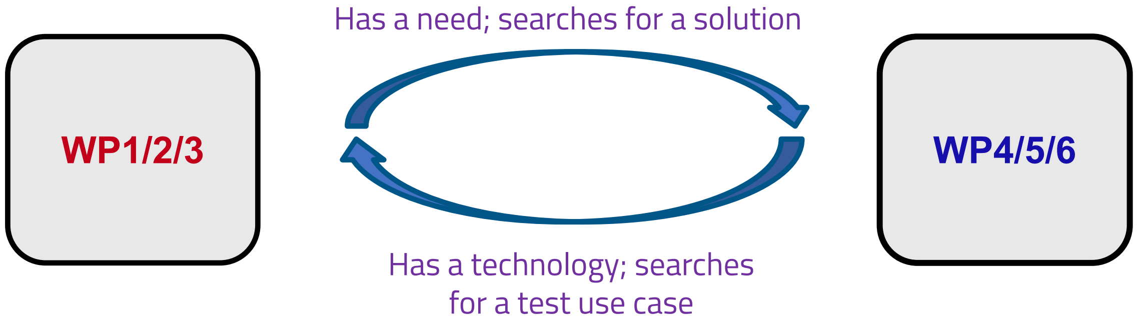
eni
INTESA SANPAOLO
LEONARDO
ThalesAlenia Space a Thales / Leonardo company
UnipolSai ASSICURAZIONI
iFAB INTERNATIONAL FOUNDATION BIG DATA AND ARTIFICIAL INTELLIGENCE FOR HUMAN DEVELOPMENT
sogei

Staff Researchers	195
(kEur)	6333
Rectruited researchers	28
(kEur)	5067
Phd positions	25
(kEur)	1992
Budget Innovation Grants (kEur)	1800
Bugdet Cascade Calls (kEur)	3200
Total Budget (kEur)	**18391**

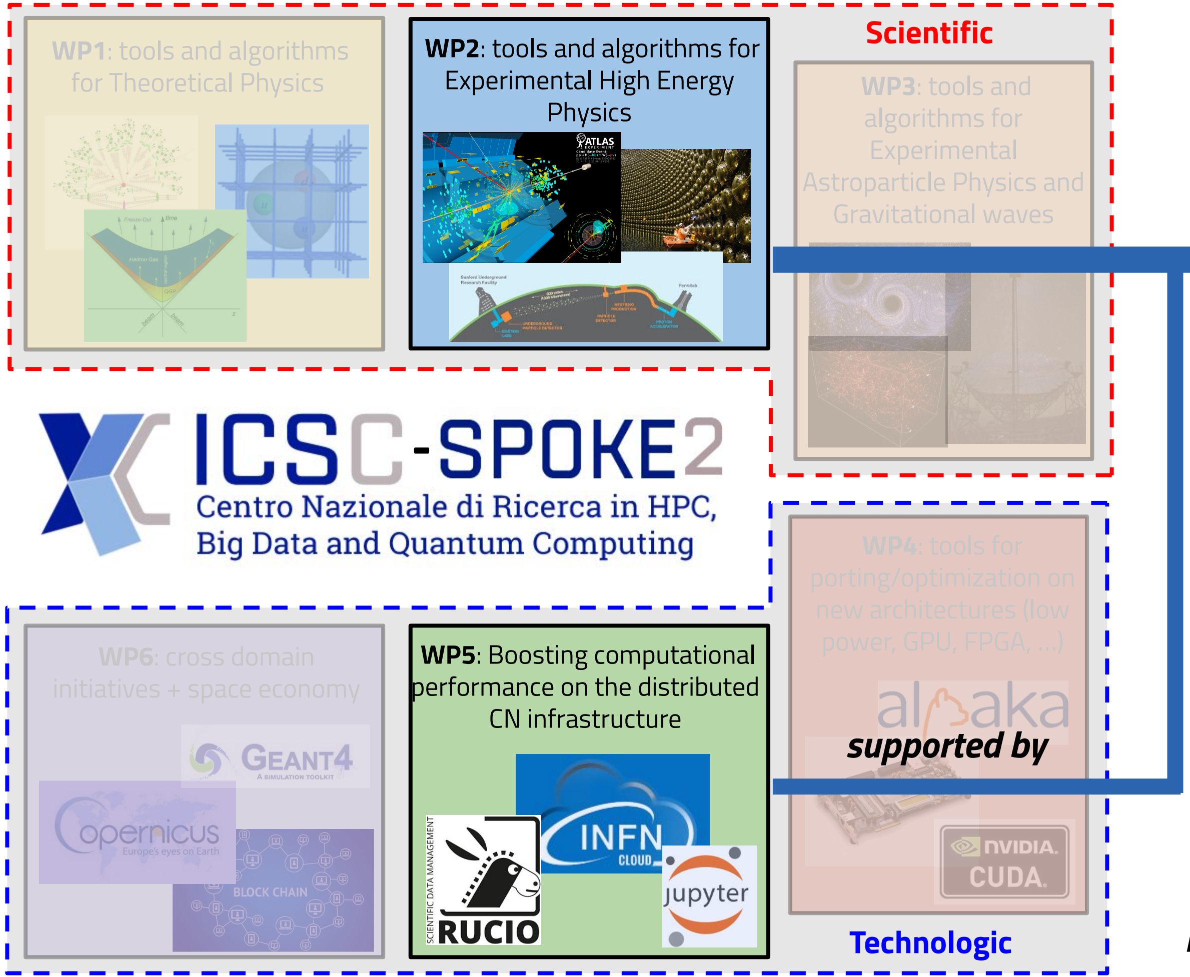
The structure of Spoke 2


Scientific
WP1: tools and algorithms for Theoretical Physics
WP2: tools and algorithms for Experimental High Energy Physics
WP3: tools and algorithms for Experimental Astroparticle Physics and Gravitational waves
ICSC-SPOKE2
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing
WP4: tools for porting/optimization on new architectures (low power, GPU, FPGA, …)
alpaka
NVIDIA. CUDA.
WP5: Boosting computational performance on the distributed CN infrastructure
INFN CLOUD
RUCIO
jupyter
WP6: cross domain initiatives + space economy
GEANT4
copernicus Europe's eyes on Earth
BLOCK CHAIN
Technologic

- We defined 2 types of Work Packages (WP):
  - **"Scientific"** WPs: they analyze the needs of the (sub-) domain, and pose open problems for which advanced computing solutions are needed;
  - **"Technological"** WPs: they harvest/investigate technical solutions in computing, on the infrastructure of the ICSC and beyond, and provide support / training for these; at the same time propose these to a larger audience, including industries.


Has a need; searches for a solution
WP1/2/3
WP4/5/6
Has a technology; searches for a test use case

Quasi interactive analysis of big data with high throughput



Scientific

WP1: tools and algorithms for Theoretical Physics

WP2: tools and algorithms for Experimental High Energy Physics

WP3: tools and algorithms for Experimental Astroparticle Physics and Gravitational waves

ICSC-SPOKE2
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

WP4: tools for porting/optimization on new architectures (low power, GPU, FPGA, …)

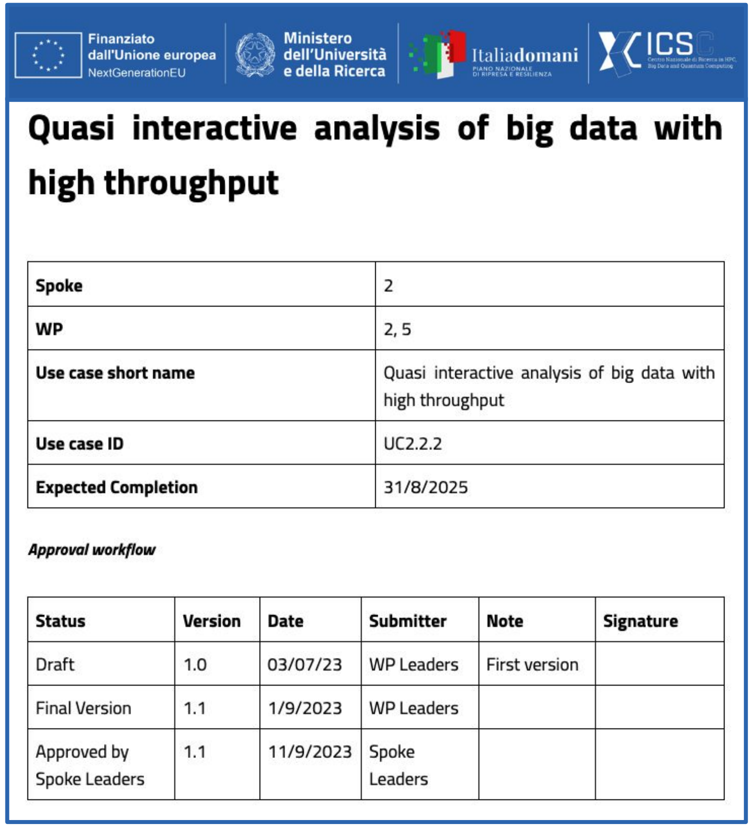
WP5: Boosting computational performance on the distributed CN infrastructure

WP6: cross domain initiatives + space economy

supported by

Technologic

ATLAS
alpaka
NVIDIA. CUDA.
GEANT4
copernicus
BLOCK CHAIN
INFN CLOUD
RUCIO
jupyter



Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani | ICSC

Quasi interactive analysis of big data with high throughput

Spoke	2
WP	2, 5
Use case short name	Quasi interactive analysis of big data with high throughput
Use case ID	UC2.2.2
Expected Completion	31/8/2025

Approval workflow

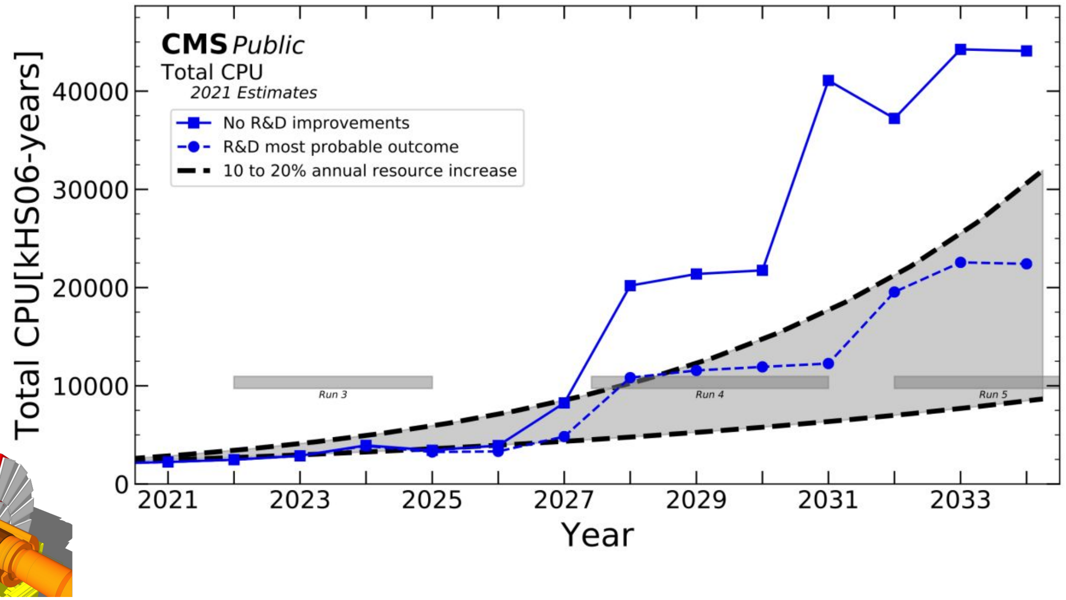
Status	**Version**	**Date**	**Submitter**	**Note**	**Signature**
Draft	1.0	03/07/23	WP Leaders	First version	
Final Version	1.1	1/9/2023	WP Leaders		
Approved by Spoke Leaders	1.1	11/9/2023	Spoke Leaders		

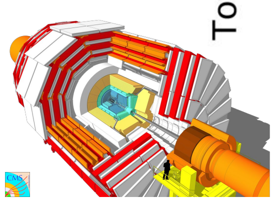
More than 40 people involved in this activity!

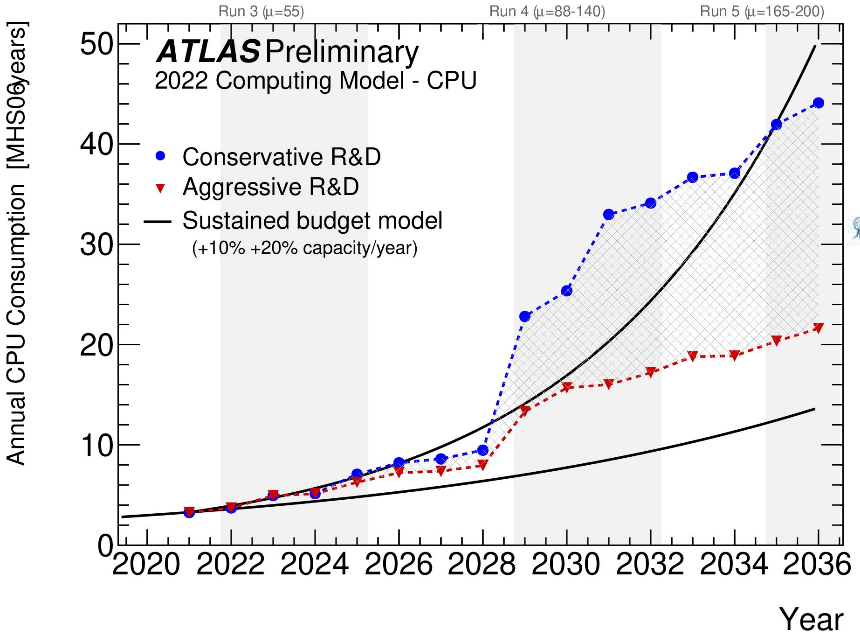
Why we need this activity?

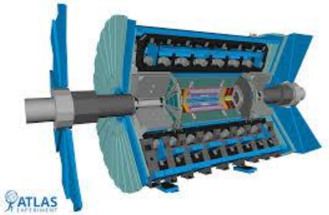
- Analysing large amounts of data efficiently, exploiting the available resources as much as possible, is a common challenge both for research and industry.

- From the beginning, the High Energy Physics (HEP) experiments at CERN, gave much attention to the computing and data management aspects. Nevertheless, the **next phases of the Large Hadron Collider** (HL-LHC) will require an even greater effort.









Why we need this activity?

Some estimate for the next 5-10 years of CMS operation:

- ~30 Billion collision events + 30 Billion simulation events;
- Each event: 2-4 kB;
- The last update of the CMS Computing model foresees this throughput:

Name	Length	% of the dataset	Data to process	Event, data rate
“A coffee”	< 5 min	1% (~0.6B evts)	~2 TB	~1.7MHz, ~7GB/s
“A lunch break”	1 hour	10% (~6B evts)	~20 TB	~1.5MHz, ~6GB/s
“A night”	12 hours	100% (60B evts)	~200 TB	~1.2MHz, ~5GB/s

- Difficult to get more than 100 Hz/CPU core → needs efficient distribution on a few tens of machines;

Not only concerning the HEP domain (“Data is data”):

- More and more scientific / industrial / societal domains have or will have soon needs similar to those from LHC:

ProtoDune: 2-3GB/s (like CMS); Real Dune: 80x

SKA: up to 2 PB/day;

CTA projects: up to 10PB/y

A single genome: ~100GB, a 1M survey=100PB

O(50 TB/y) per sensor; ~10-100 sensors: O(5 PB/y)

New analysis paradigm based on:

- Declarative programming and interactive workflows;
- Distributed computing on geographically separated resources.

Why we need this activity?

Some estimate for the next 5-10 years of CMS operation:

- ~30 Billion collision events + 30 Billion simulation events;
- Each event: 2-4 kB;
- The last update of the CMS Computing model foresees this throughput:

Name	Length	% of the dataset	Data to process	Event, data rate
"A coffee"	< 5 min	1% (~0.6B evts)	~2 TB	~1.7MHz, ~7GB/s
"A lunch break"	1 hour	10% (~6B evts)	~20 TB	~1.5MHz, ~6GB/s
"A night"	12 hours	100% (60B evts)	~200 TB	~1.2MHz, ~5GB/s

- Difficult to get more than 100 Hz/CPU core → needs efficient distribution on a few tens of machines;

Not only concerning the HEP domain ("Data is data"):

- More and more scientific / industrial / societal domains have or will have soon needs similar to those from LHC:

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

ProtoDune: 2-3GB/s (like CMS); Real Dune: 80x

SKA SQUARE KILOMETRE ARRAY

SKA: up to 2 PB/day;

CTAO

CTA projects: up to 10PB/y

A single genome: ~100GB, a 1M survey=100PB

O(50 TB/y) per sensor; ~10-100 sensors: O(5 PB/y)

High Throughput Platform

New analysis paradigm based on:

- Declarative programming and interactive workflows;
- Distributed computing on geographically separated resources.

Re-thinking the analysis pipeline

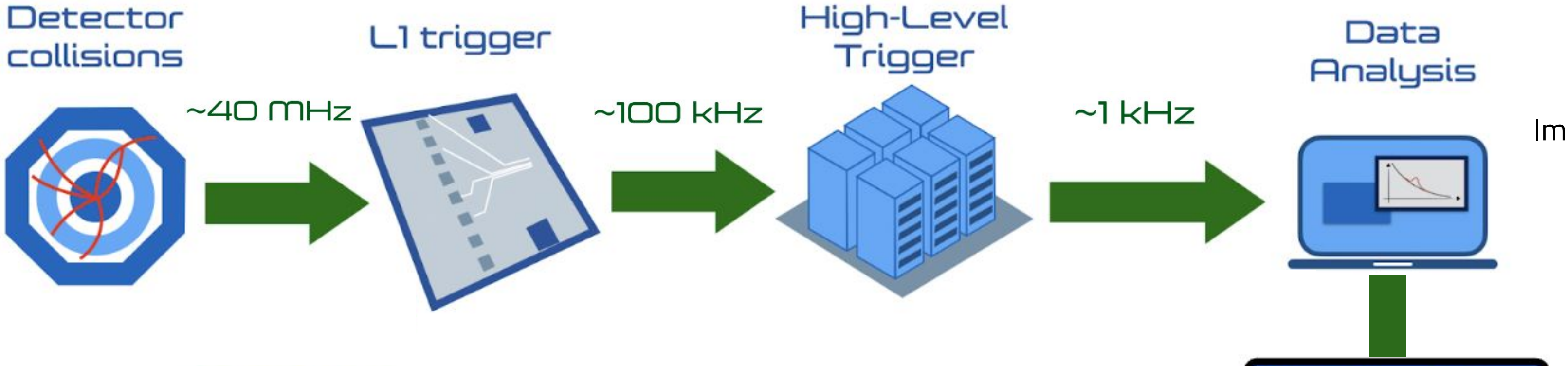
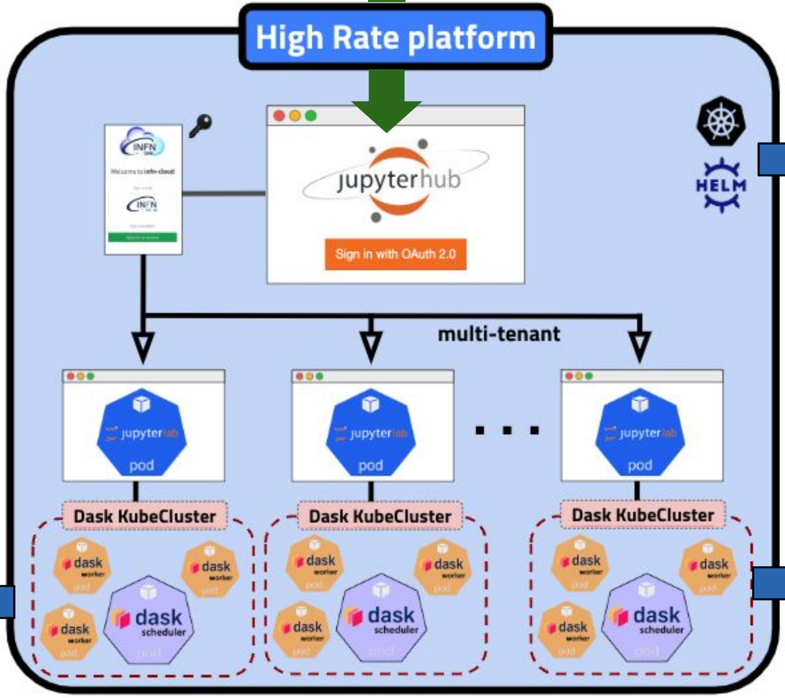


Image taken from: link

More on T.Tedeschi talk



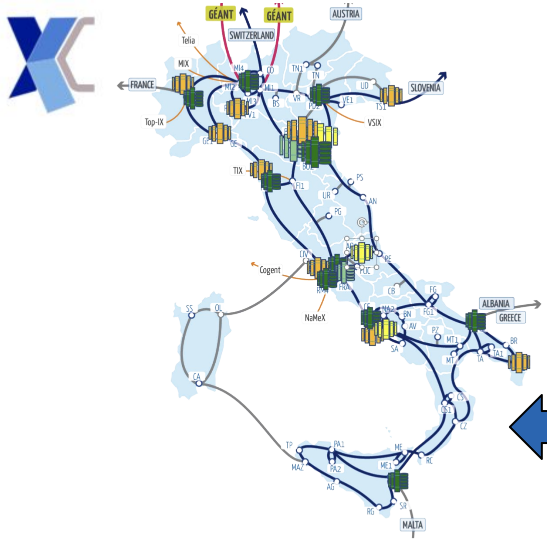
Deployment of the Kubernetes resources handled via HELM Charts.

↓

(Scalable on available resources)

- The execution happens in the Dask Cluster;
- Users choose on how many cores parallelly distribute the analysis.

offloaded to...



Activities (so far) orbiting around the flagship

Vector Boson Scattering ssWW analysis in hadronic tau and light lepton

Heavy Neutral Lepton search on heavy neutrinos in the D_S decays

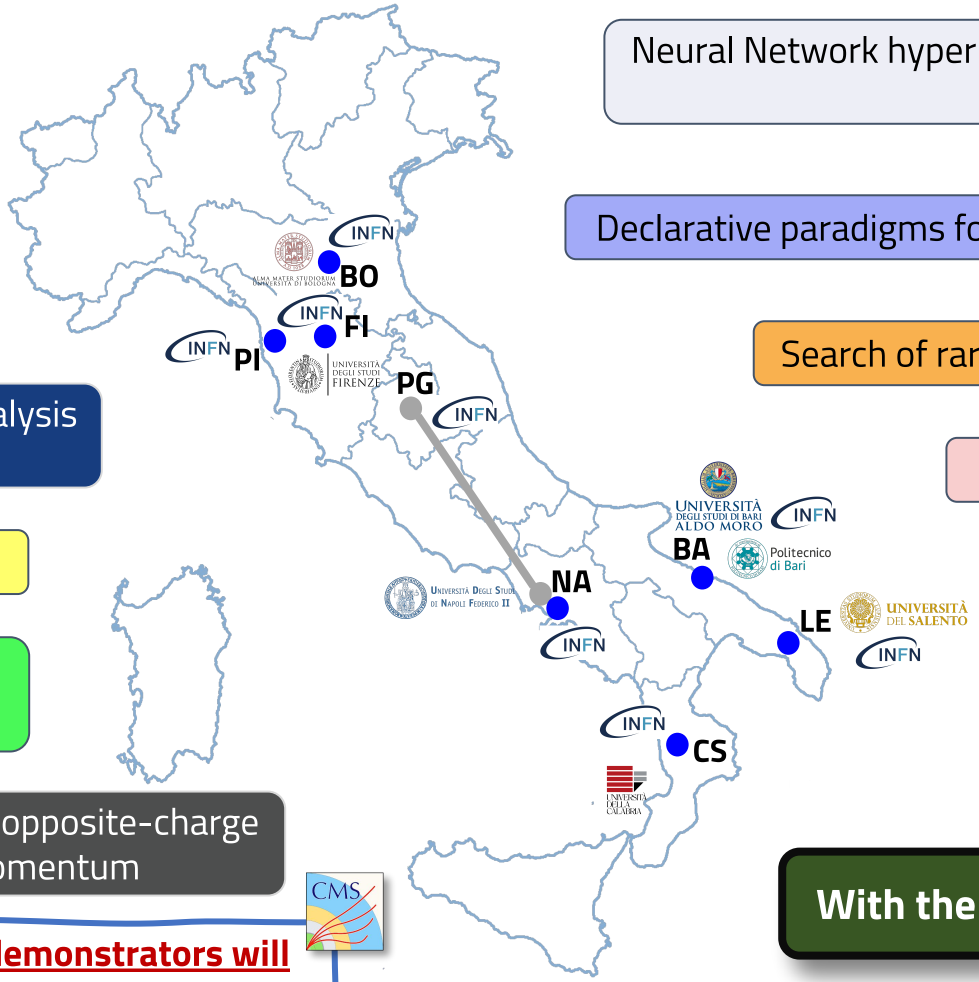
Muon detector performance analysis

Continuous Integration pipeline, triggering analysis execution on HTP

di-Higgs decaying to two b quarks and two muons

Differential cross section measurement for ttbar inclusive production

Search for new phenomena in events with two opposite-charge leptons, jets and missing transverse momentum


INFN
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
BO
INFN
FI
INFN
PI
UNIVERSITÀ DEGLI STUDI FIRENZE
PG
INFN
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
INFN
BA
Politecnico di Bari
Università degli Studi di Napoli Federico II
NA
INFN
LE
UNIVERSITÀ DEL SALENTO
INFN
INFN
CS
UNIVERSITÀ DELLA CALABRIA
CMS

Specific analyses demonstrators will be shown on the closed CMS session!

Neural Network hyperparameter optimisation applied to future colliders (FCC-ee)

Declarative paradigms for analysis description and implementation

Search of rare events in tau to 3 muons decay

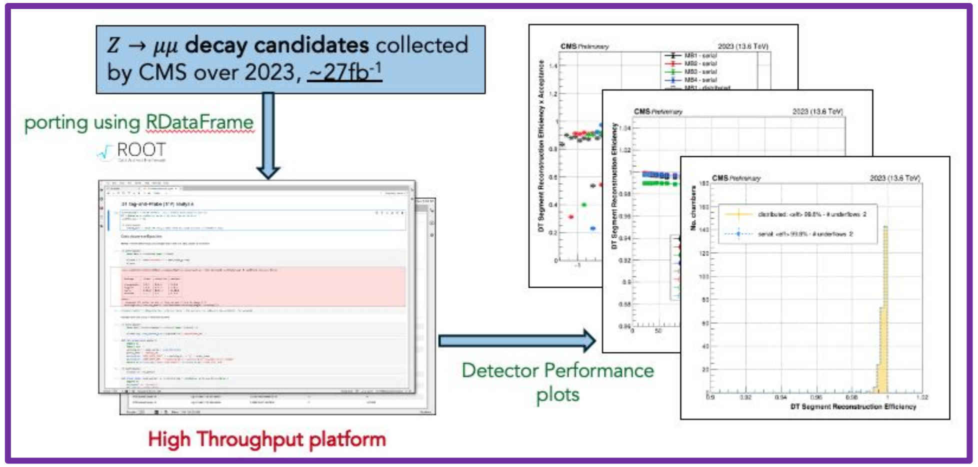
top quark+MET analysis

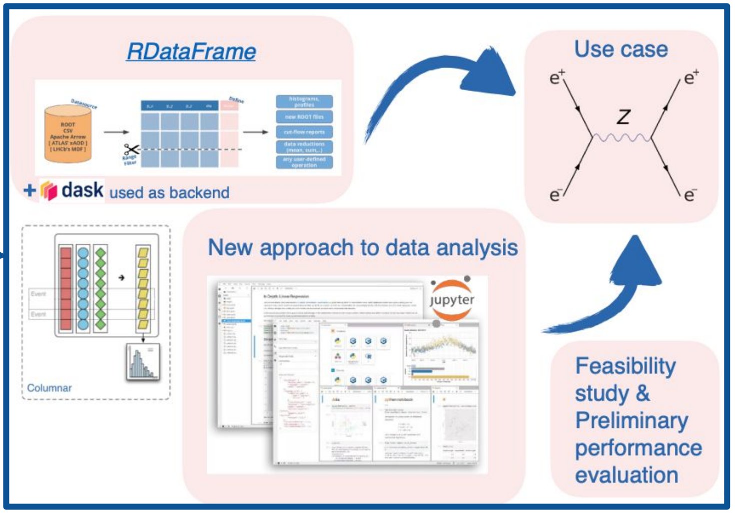
Benchmark interactive analysis for future colliders (FCC-ee)

With the infrastructural support of WP5

Scientific production in conferences

- Poster at the "International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT 2024)":
  - *Declarative paradigms for analysis description and implementation.*
  - *Quasi interactive analysis of High Energy Physics big data with high throughput.*
- Talk at the "Incontri di Fisica delle Alte Energie (IFAE 2024)":
  Analisi quasi-interattiva per big data con alto throughput per la Fisica delle Alte Energie.
- Talk at the "International Conference on High Energy Physics (ICHEP 2024)":
  Enhancing CMS data analyses using a distributed high throughput platform.
- Talk at the "2nd European Committee for Future Accelerator (ECFA) Workshop on Higgs/EW/Top Factories":
  Benchmark interactive analysis for future colliders.
- Talk at the "Conference on Computing in High Energy and Nuclear Physics (CHEP 2024)":
  Leveraging distributed resources through high throughput analysis platforms for enhancing HEP data analyses.


$Z \to \mu\mu$ **decay candidates** collected by CMS over 2023, ~27fb^{-1}
porting using RDataFrame
ROOT
Detector Performance plots
High Throughput platform
CMS Preliminary
2023 (13.6 TeV)
DT Segment Reconstruction Efficiency


RDataFrame
+ dask used as backend
Use case
e^+
e^+
Z
e^-
e^-
New approach to data analysis
jupyter
Columnar
Feasibility study & Preliminary performance evaluation

Why this workshop?

- The challenge presented by the next LHC phases requires a strong development effort of new tools, for making data analysis as efficient and as modern as possible;
- Several analysis from the HEP world are already testing such infrastructure, for performance measurements;
- Thanks to the ICSC national center, we have the unique opportunity to build a modern infrastructure for research, aligned with the needs from High Energy Physics and beyond;
- What we need, at this point, is to form scientists (young and senior), to exploit and perform their research activities in the most efficient and modern way possible!
    - This is our motivation for this event (**hoping for more to come**)!

For now…

Enjoy the workshop!