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

Leveraging distributed resources through high throughput analysis platforms for enhancing HEP data analyses

Adelina D'Onofrio[1], Tommaso Diotalevi[1,3], Francesco Giuseppe Gravili[1,5], Salvatore Loffredo[1,2], Elvira Rossi[1,2], Federica Maria Simone[1,4], Bernardino Spisso[1]
on behalf of the ATLAS and CMS Collaborations

1 INFN, 2 University Federico II, 3 University of Bologna, 4 Polytechnic Bari, 5 Università del Salento

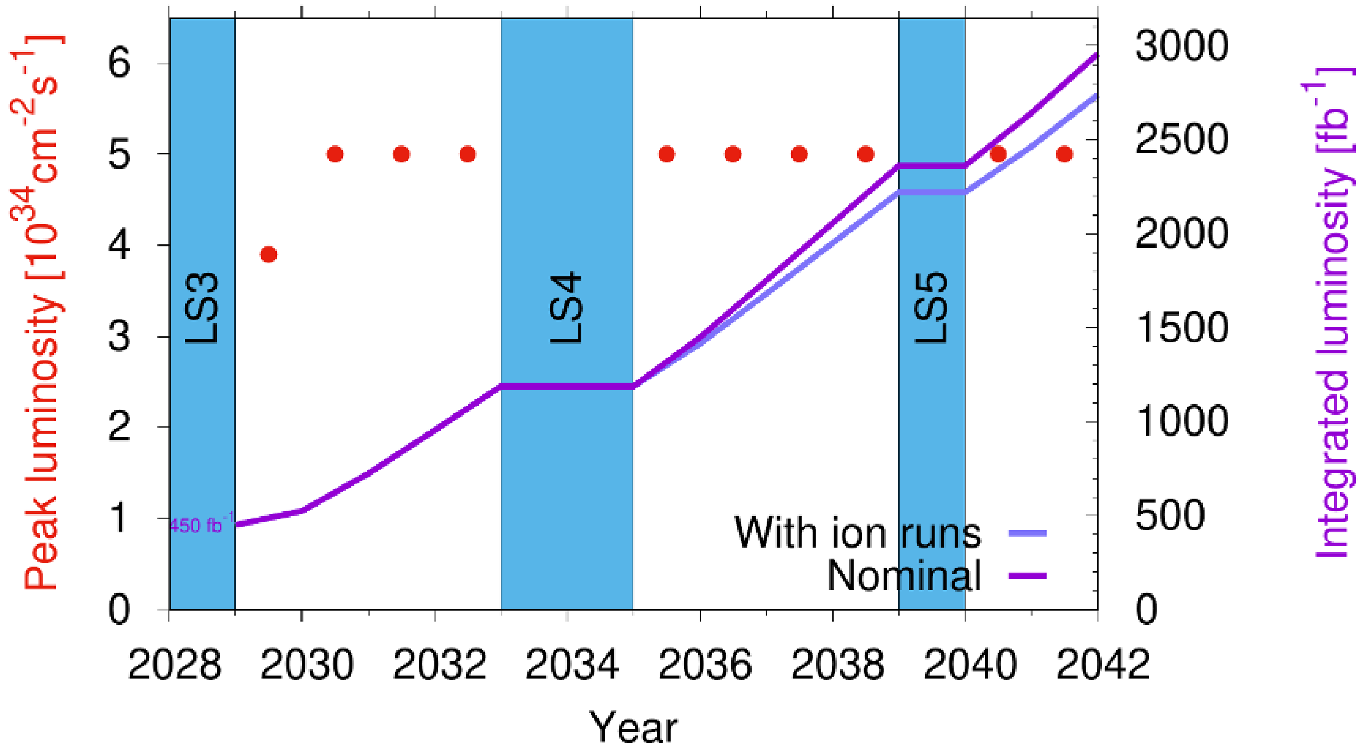
On behalf of the ATLAS and CMS Collaborations
CHEP2024, 19-25 Oct 2024, Krakow

Outline

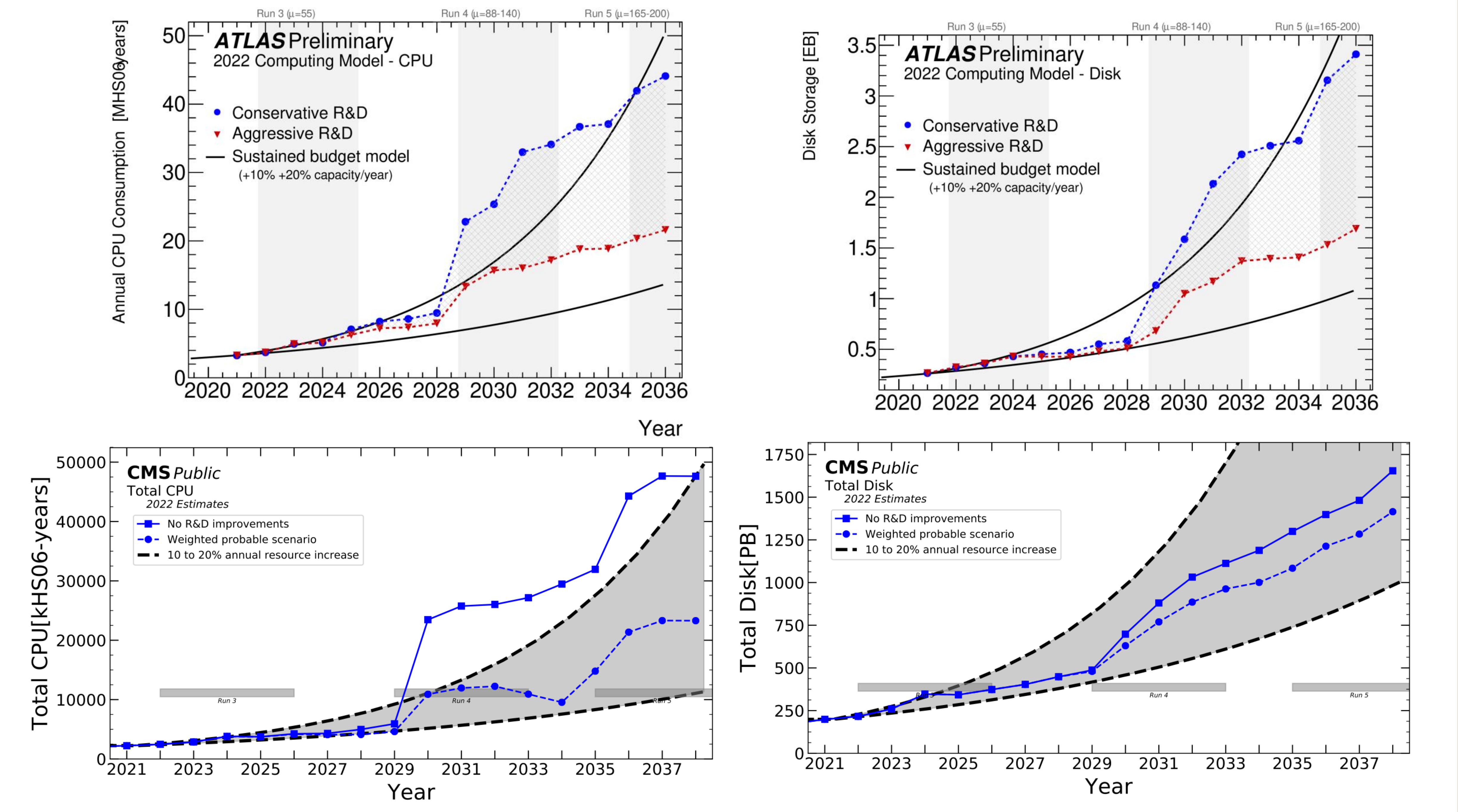
- Motivations
- Test infrastructure
- Analysis use-cases:
  - Search for SUSY signatures at ATLAS
  - Flavor physics and search for rare decays at CMS
    - Preliminary scalability results
- Conclusions

Motivations

- **Challenges of LHC, and HL-LHC** are pushing to **re-think the HEP computing models**
  - Impact on several aspects, from software to the computing infrastructure



Similar trends for ATLAS and CMS HL-LHC projections

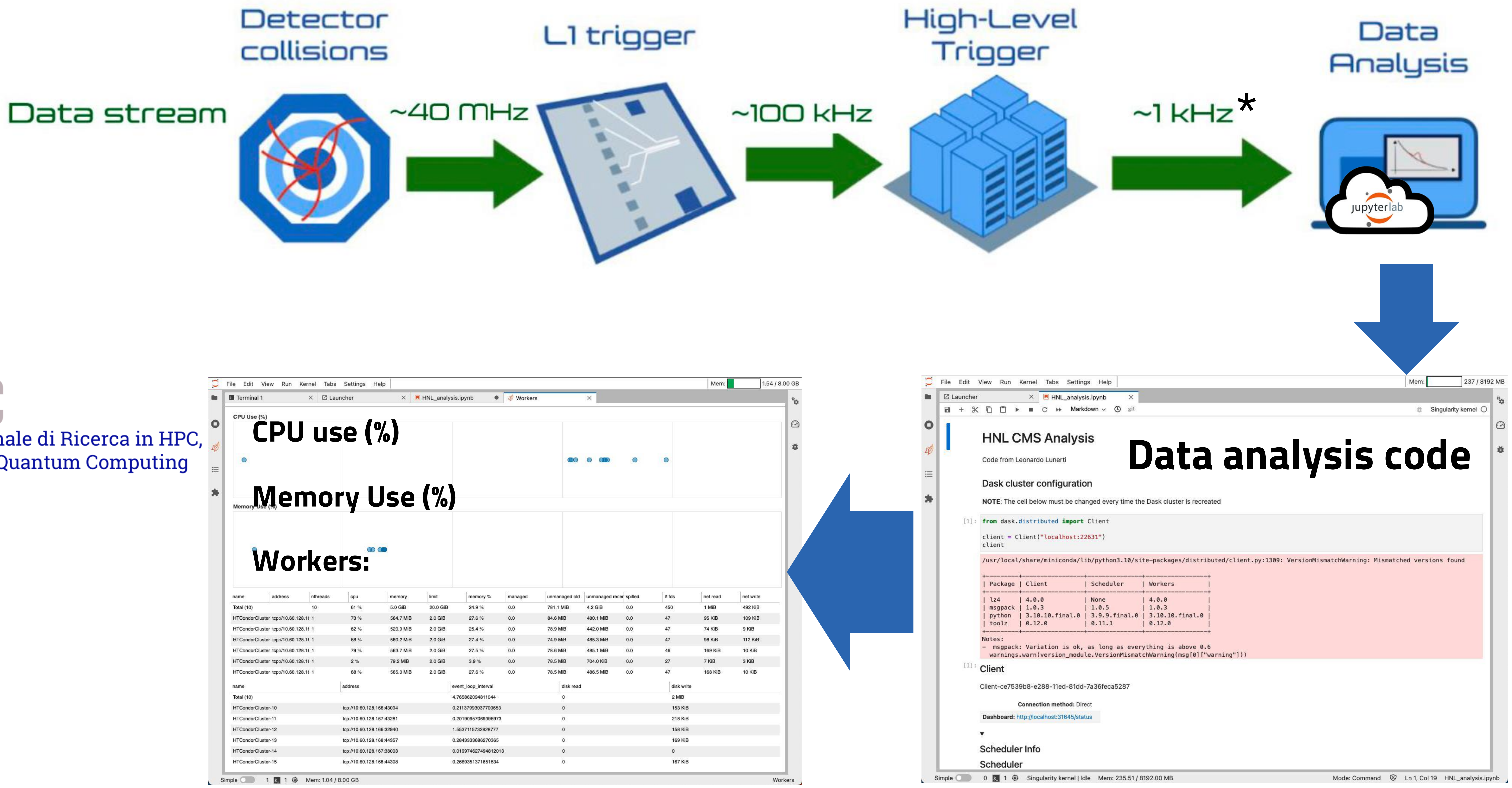


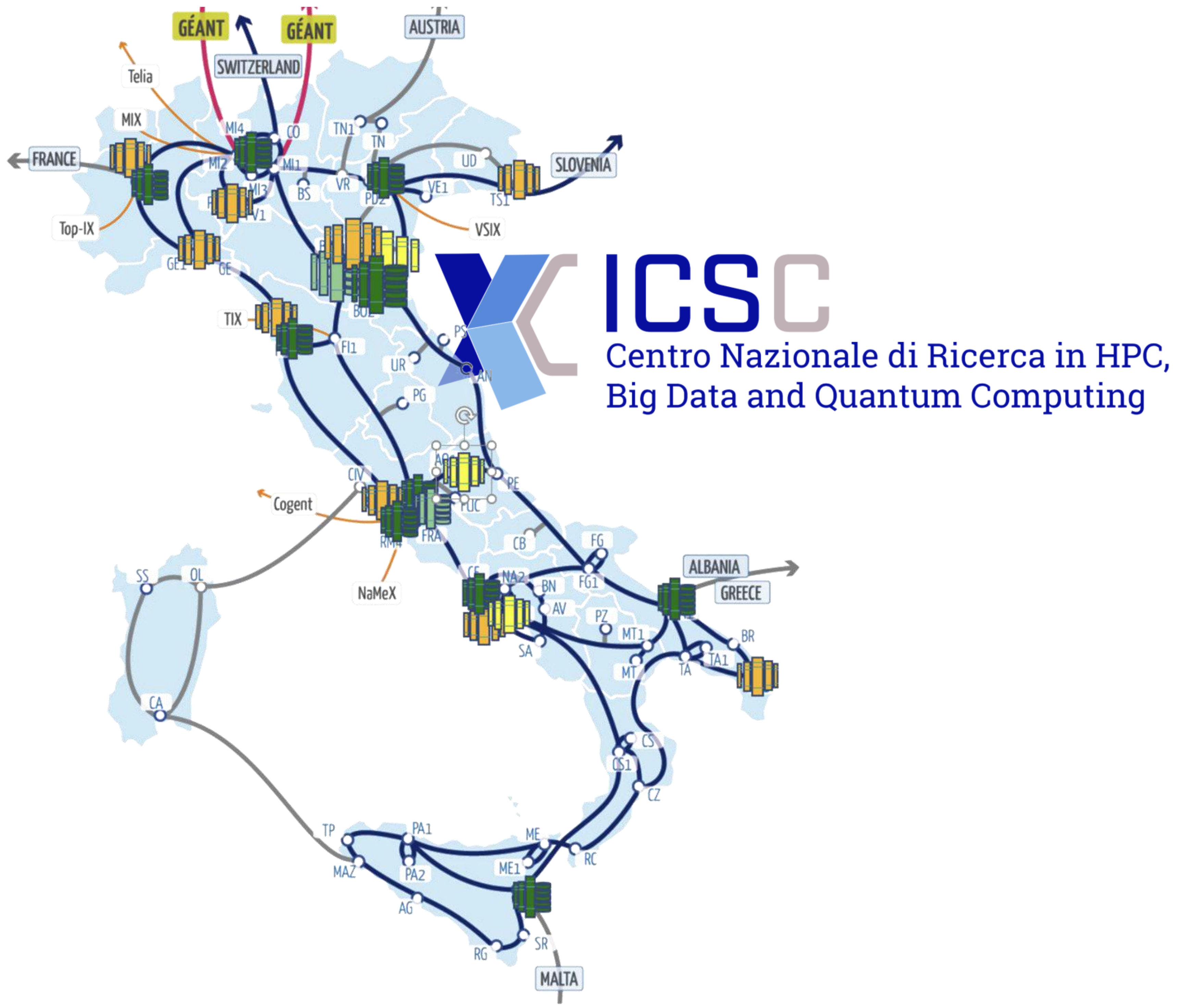
Higher rates of collision events → **Higher demand for computing and storage resources**

Need to:

- Optimize the usage of CPU and storage
- Promote the usage of better data formats
- **Develop new analysis paradigms**!
- New software based on declarative programming and interactive workflows
- Distribute on geographically separated resources

HEP data analysis with ICSC



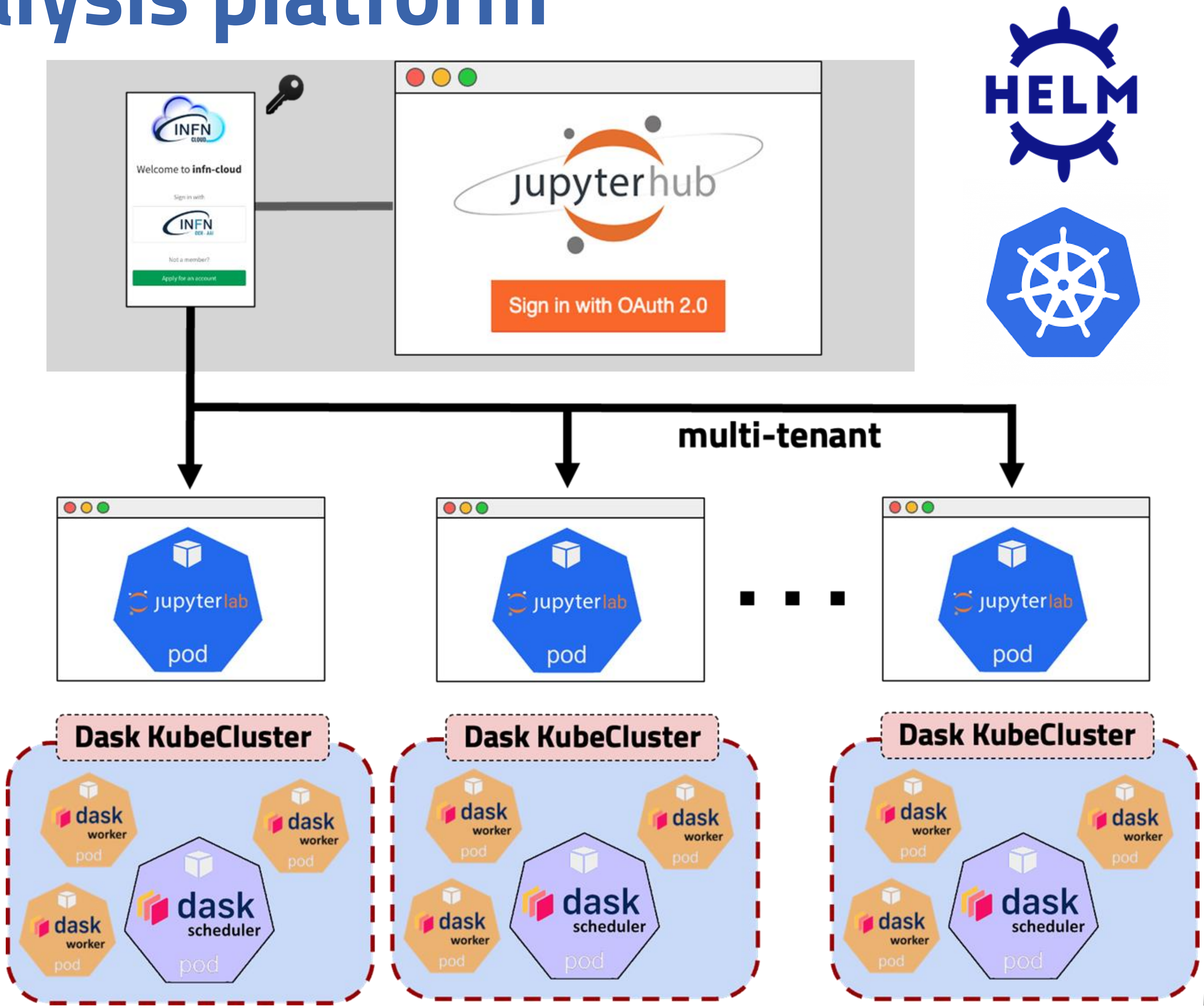


*trigger rates for previous Runs, now factor 3 ÷ 5 higher, will further scale in HL-LHC

High throughput data analysis platform

- After connecting to an entrypoint URL, the user reaches a Jupyterhub instance that, after authentication and authorization via INDIGO-IAM, allocates the required resources for the user's working area.

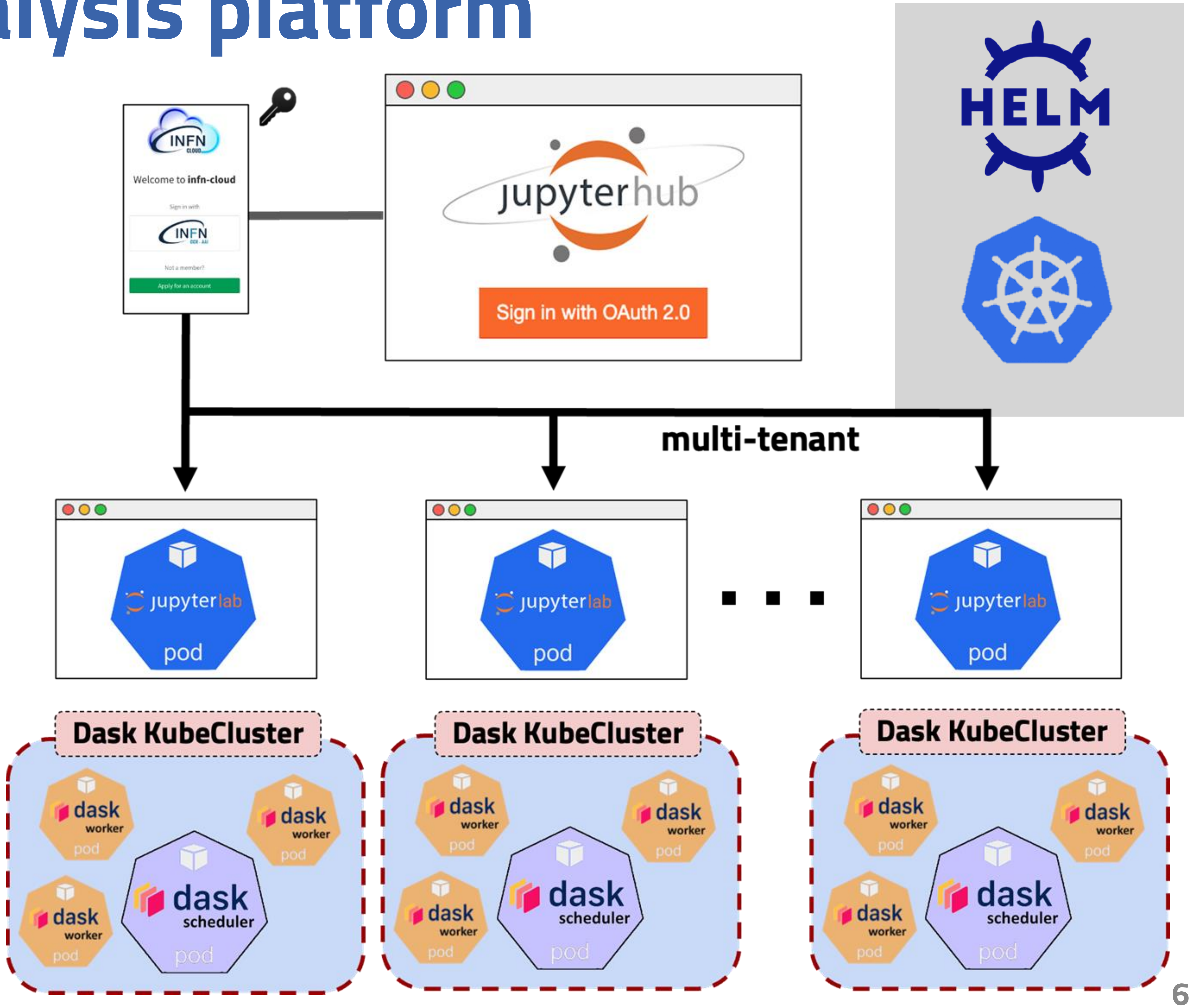
- The jupyterhub is deployed on a Kubernetes (k8s) cluster with **128 vCPUs and 258 GB**, divided into 8 nodes configured via RKE2



High throughput data analysis platform

- The deployment of the Kubernetes resources is handled via HELM charts in the official Spoke2 Jhub HELM repo

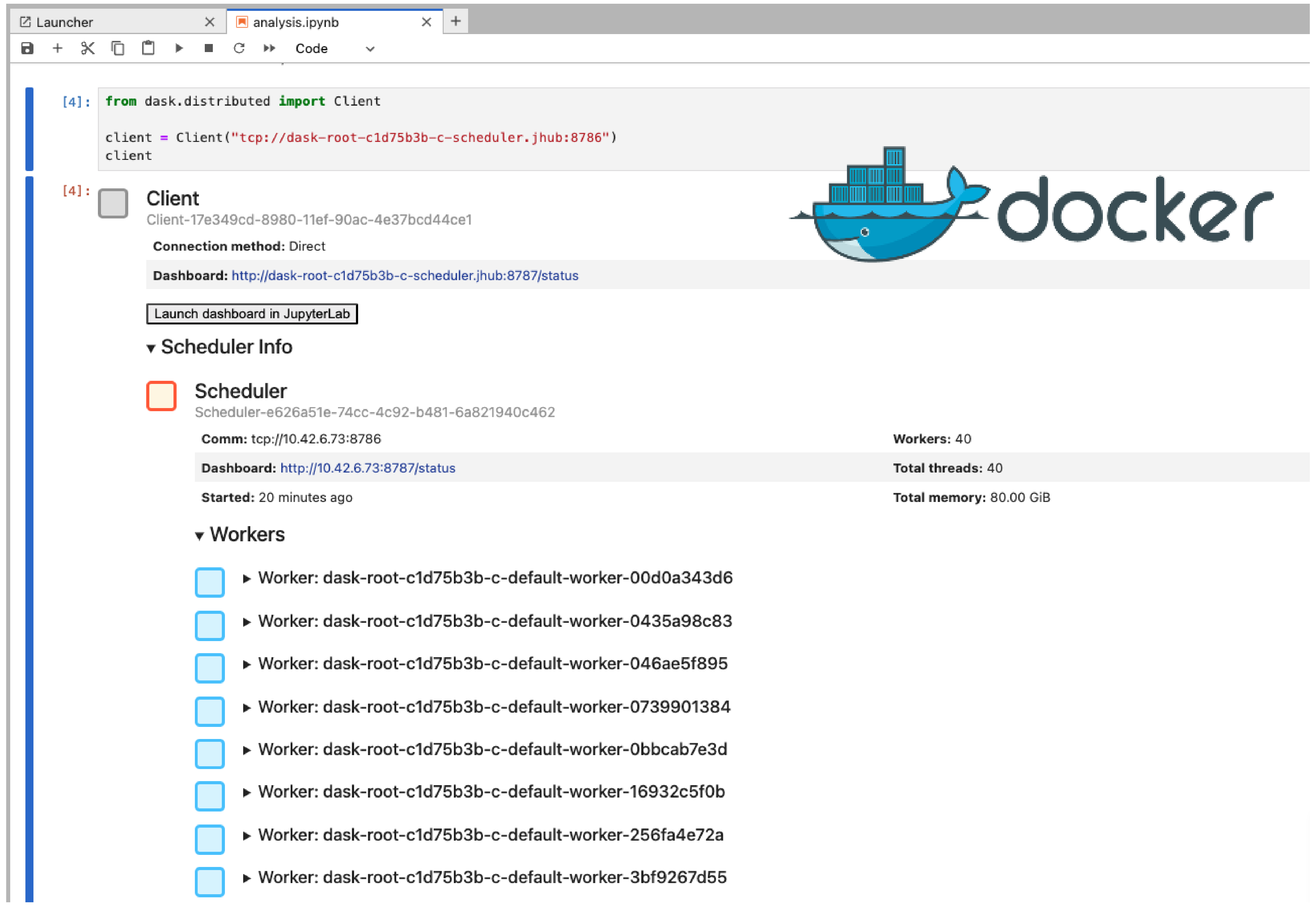
- This allows for a scalable and fault-tolerant deployment of the available resources

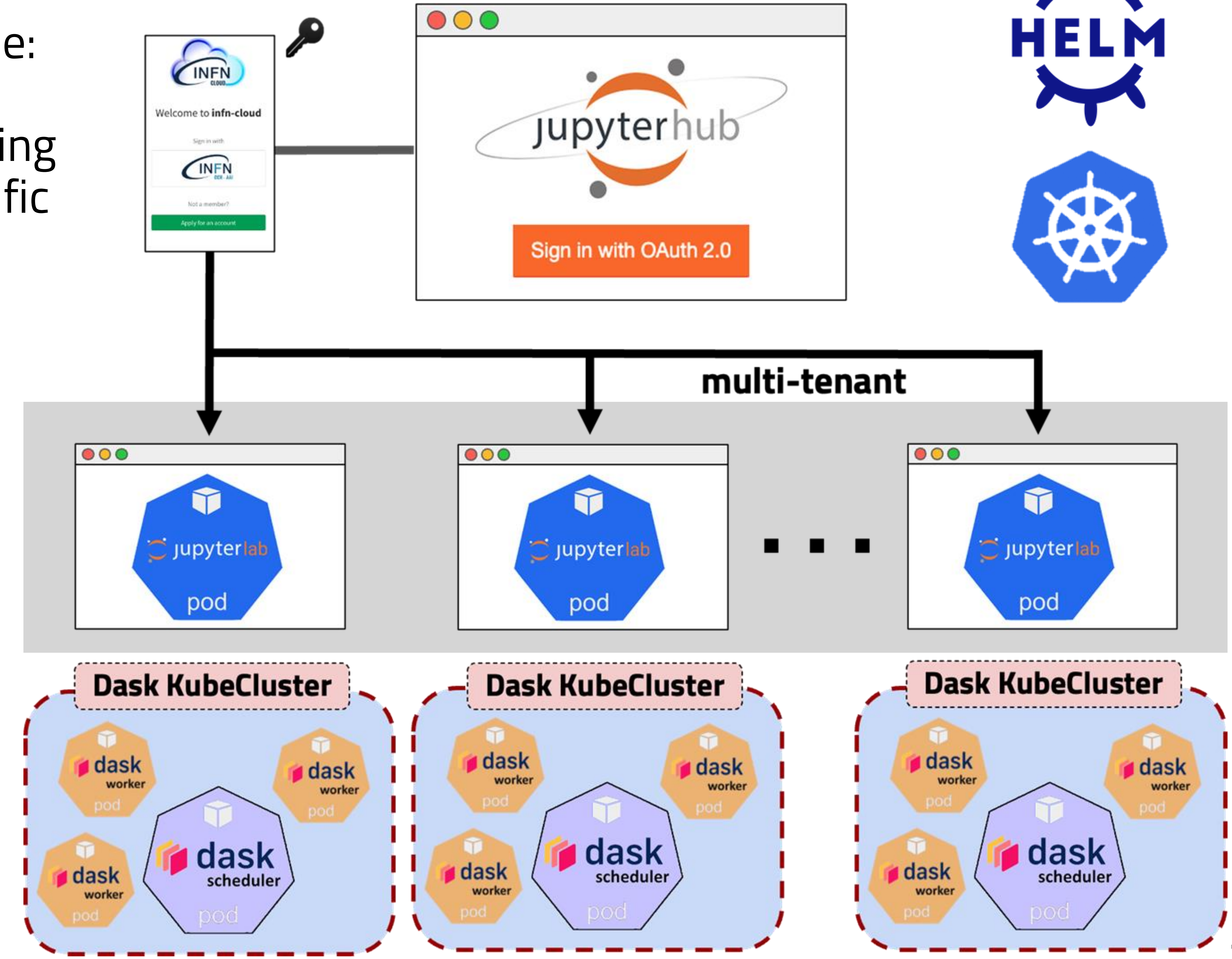


Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

High throughput data analysis platform

- Jupyterlab interface is flexible and customizable:
  - Includes specific plugins (e.g. Dask)
  - Working environment highly customizable using Docker containers allowing for experiment specific software



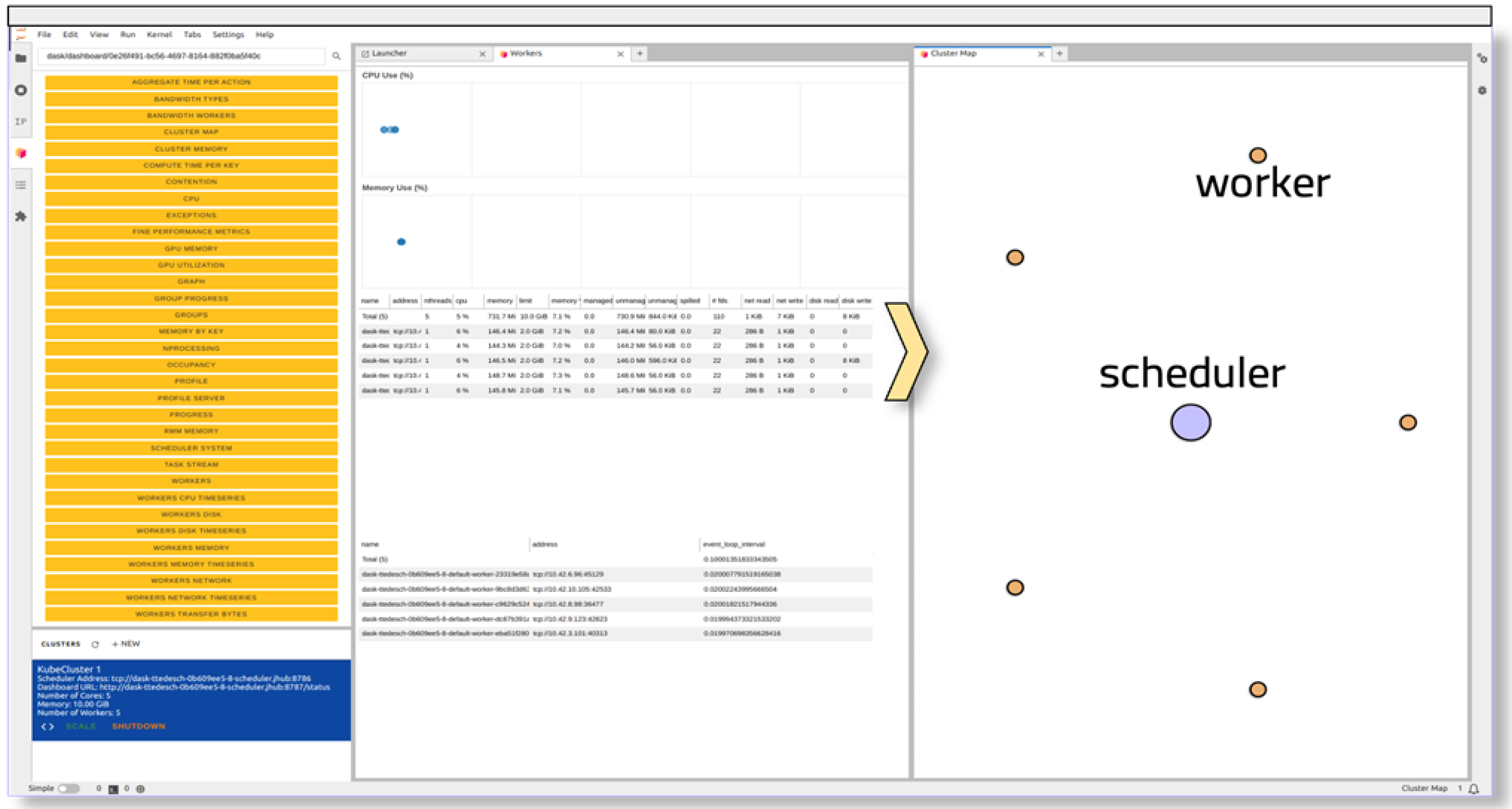


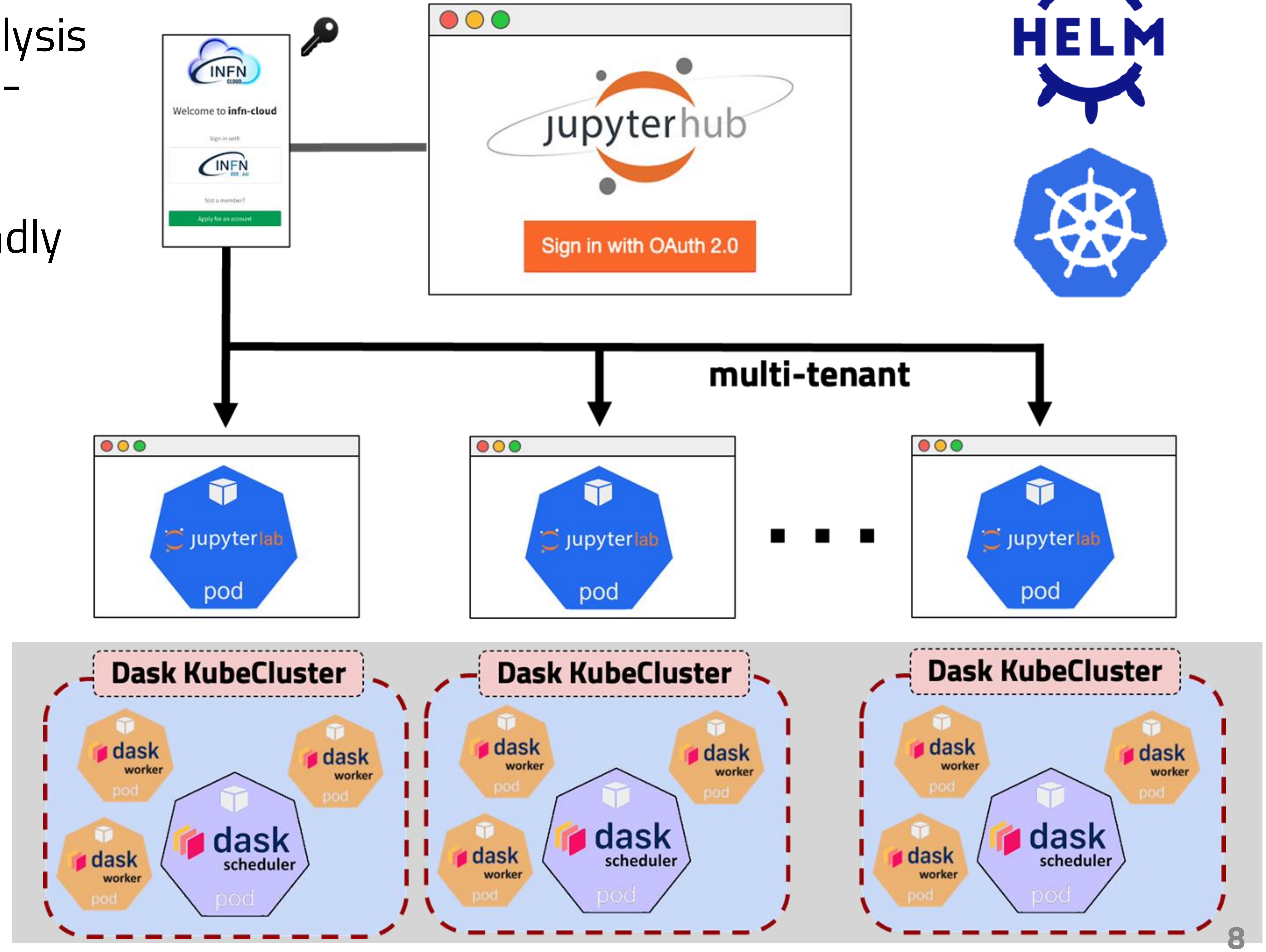
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

High throughput data analysis platform

- Ideal environment for testing interactive analysis and validating new frameworks, e.g. the multi-threading features of ROOT RDataFrame
- The Dask Labextension provides a user-friendly monitoring dashboard
- More in the official docs!

Dask Dashboard 1 **Monitoring workers** 2 **Cluster map**

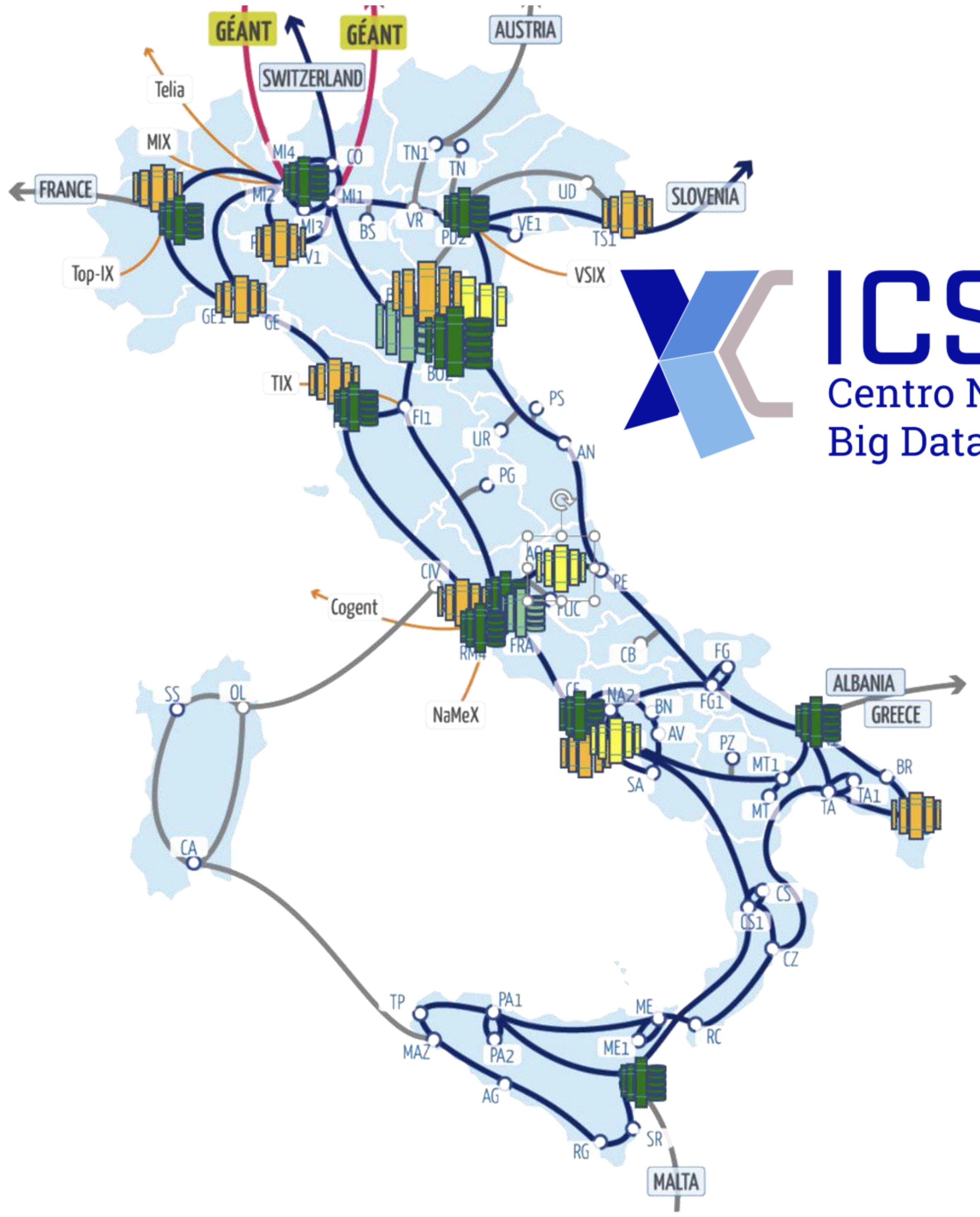




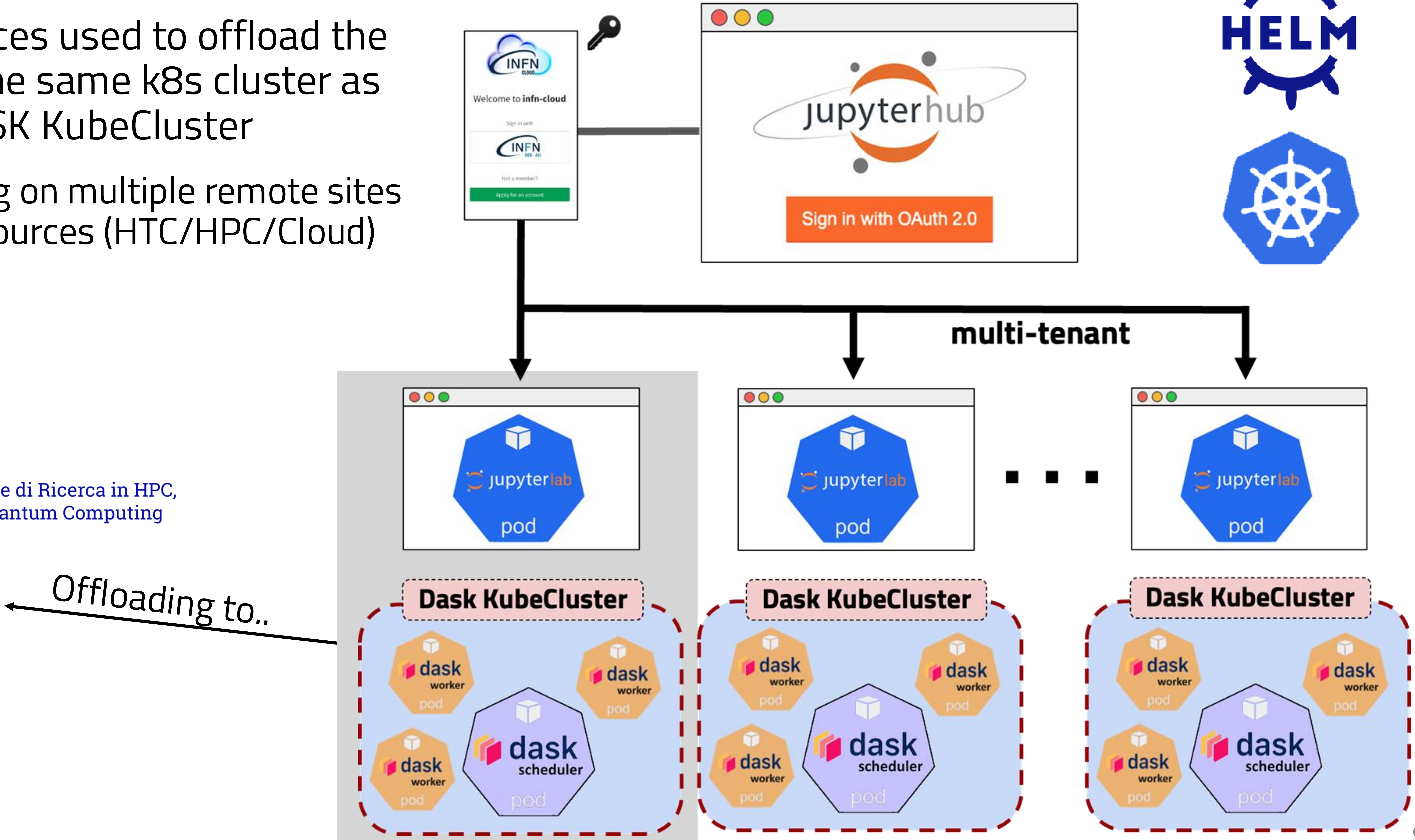
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

High throughput data analysis platform

- Offloading strategy: resources used to offload the computation are hosted in the same k8s cluster as the jupyter interface, via DASK KubeCluster
- **Under development:** spawning on multiple remote sites allowing for heterogeneous resources (HTC/HPC/Cloud) (see more in backup)



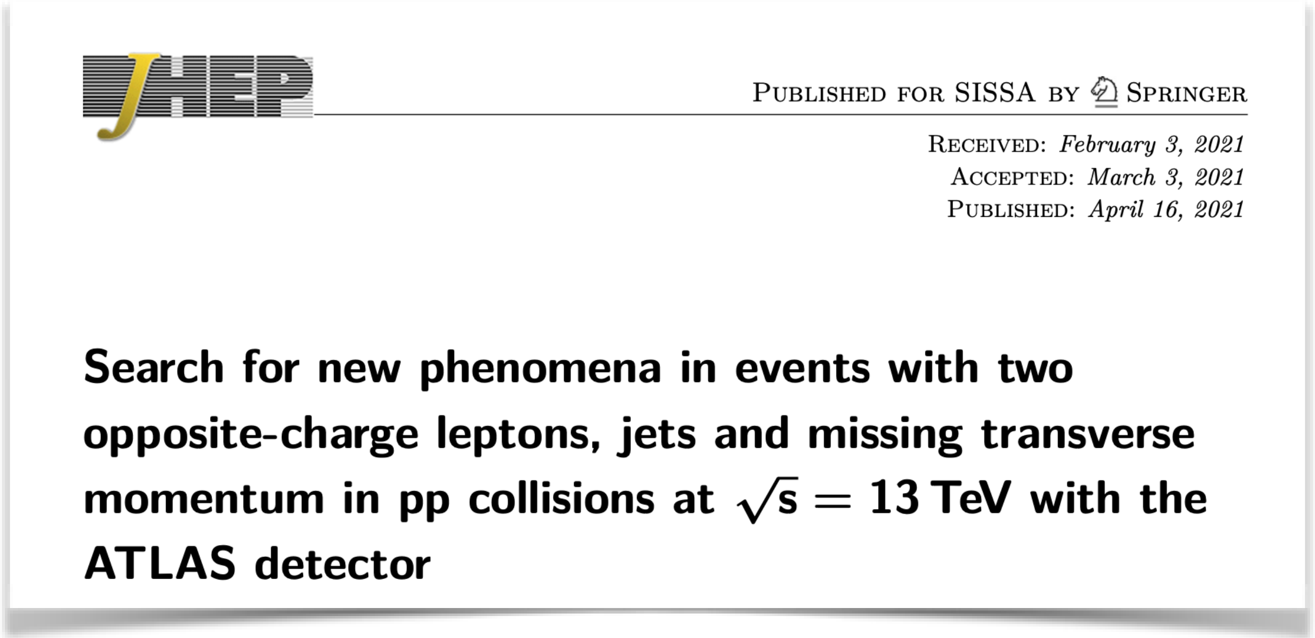
ICSC
Centro Nazionale di Ricerca in HPC,
Big Data and Quantum Computing



Benchmark interactive analyses

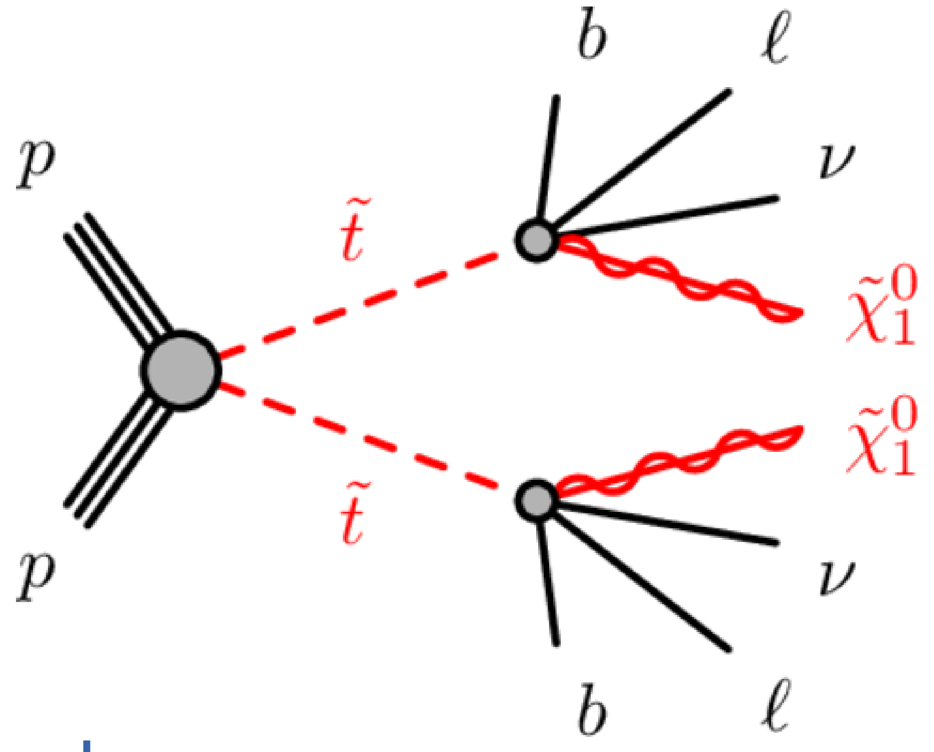
ATLAS use-case

SUperSYmmetry: Beyond Standard Model (BSM) theory


JHEP

Published for SISSA by Springer
Received: February 3, 2021
Accepted: March 3, 2021
Published: April 16, 2021

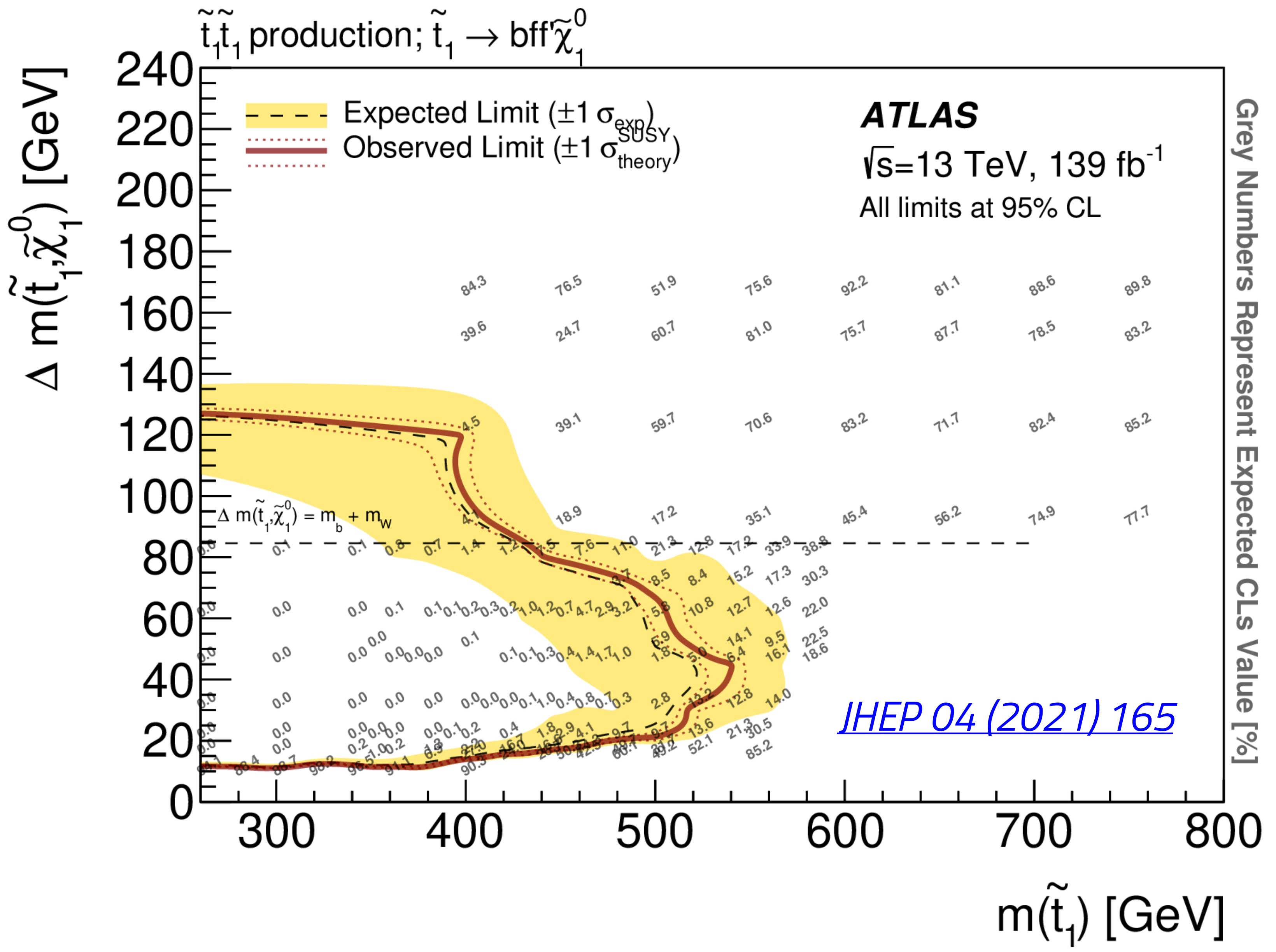
Search for new phenomena in events with two opposite-charge leptons, jets and missing transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector



Soft leptons coming from a virtual W* boson decay

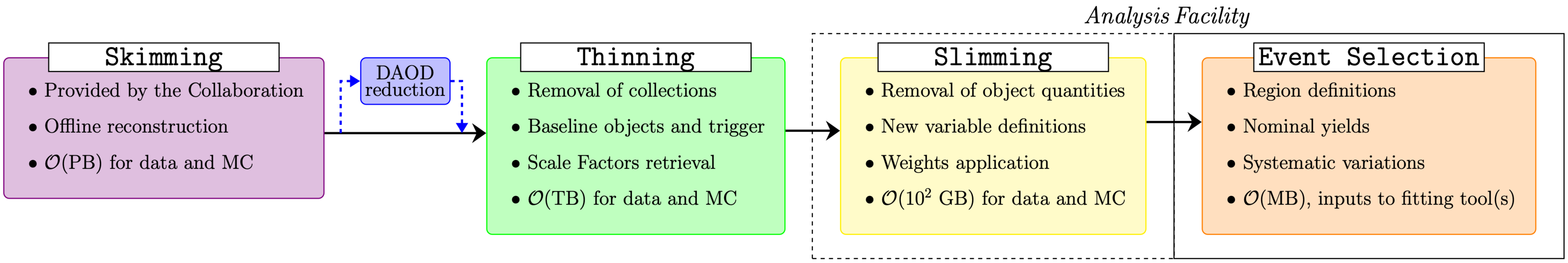
Compressed mass spectra: $\Delta m < m_W + m_b$

- Three different analysis in the ***Run 2 paper***, already published, according to mass splitting between *stop* ($\tilde{t}_1$) and *neutralino* ($\tilde{\chi}^0{}_1$), allowing different decay modes:
  - 2 body → $\Delta m > m_t$
  - 3 body → $m_W + m_b < \Delta m < m_t$
  - 4 body → $\Delta m < m_W + m_b$ **used as a benchmark**
- Common final state signature: 2 OS leptons from W* decays, b-jets and missing transverse energy
- Cut-based analysis



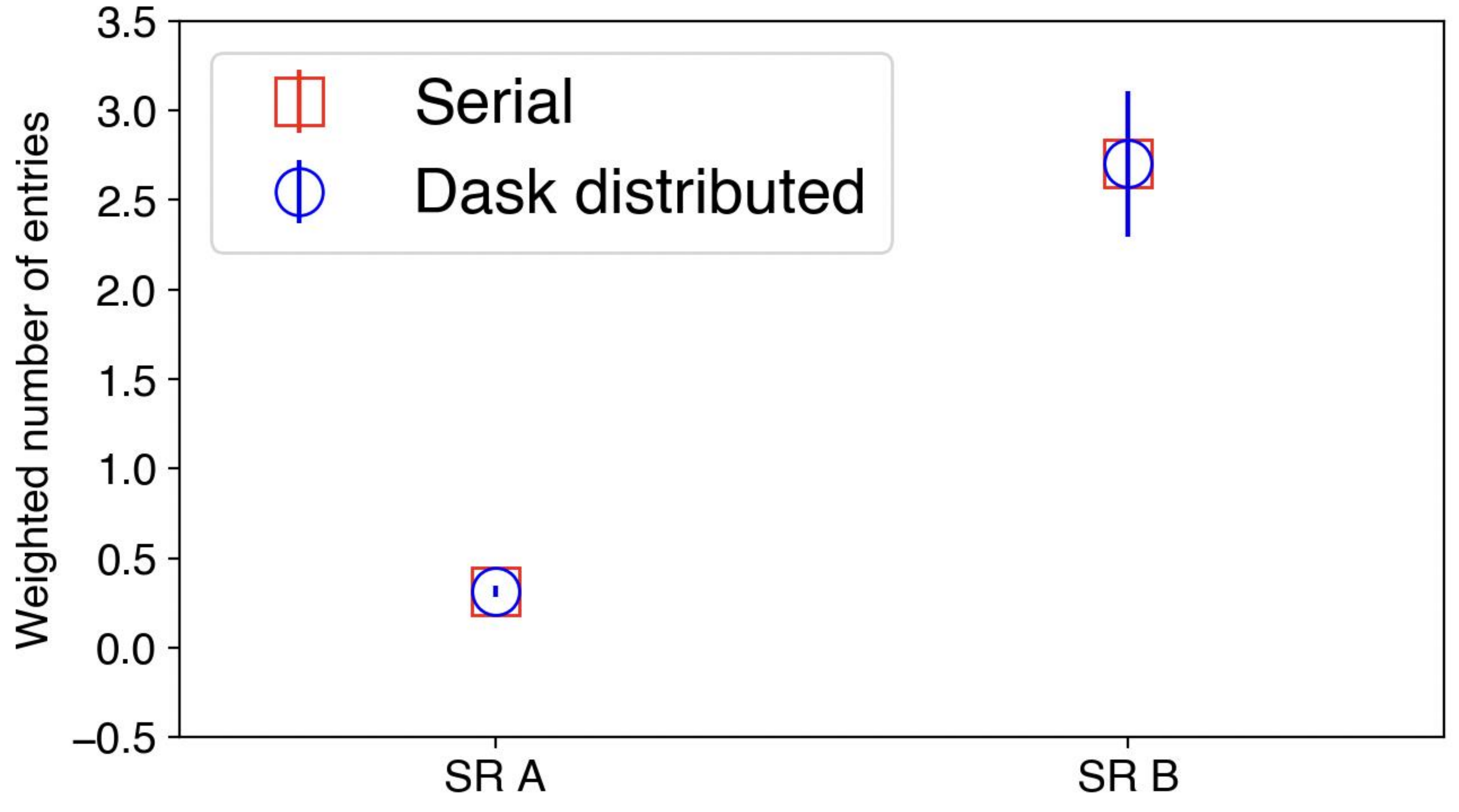
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

4-body search workflow



Sanity check

Weighted number of events in the Wt background sample, after the event selection cuts in signal regions A and B, nominal case →identical



Slimming

ATLAS slimming code already in RDataFrame, but entirely written and compiled in C++ —> Dask distributed approach was not used

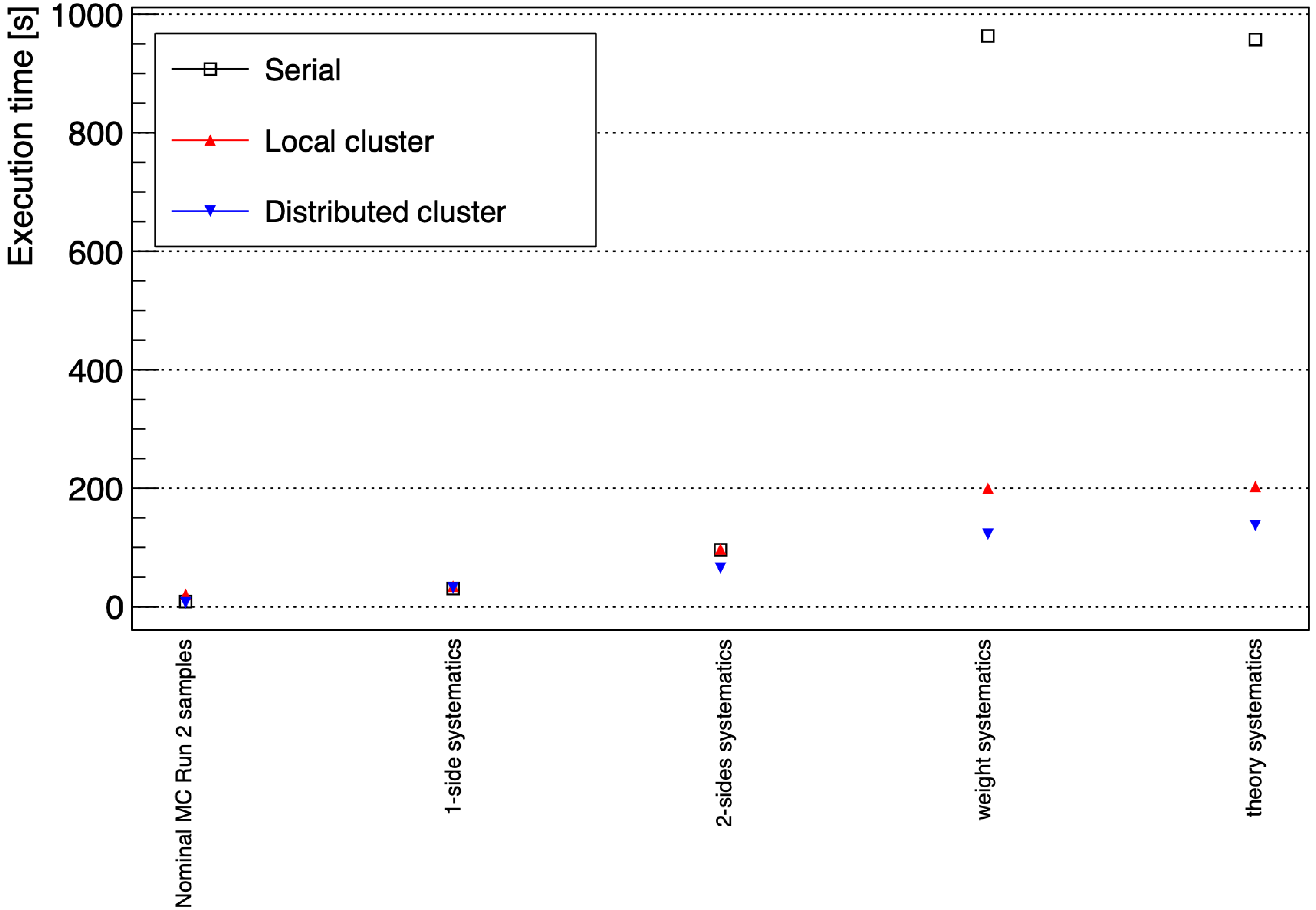
Event Selection

- Event selection for fitting tools
- RDataFrame + Dask applied to Wt bkg sample ~ 1.8 GB copied to the INFN workspace
- Playing with syst. variations
- Code ready to play with other backgrounds

Preliminary results

Defined Metric	
Overall execution time	Time elapsed from the start of the execution (execution triggered) to the end of execution

- Exploiting the distributed approach, the execution time improves *wrt* the standard/serial approach if we iterate over a significative number of systematic variations (each step in the x-axis includes previous contributions)



- LocalCluster: Dask multithread execution on local machine (max 8 cores, 16 GiB)
- Distributed: Dask distributed execution on remote workers

Scheduler and Working Nodes Reports

Distributed approach

Dask Performance Report

Select different tabs on the top for additional information

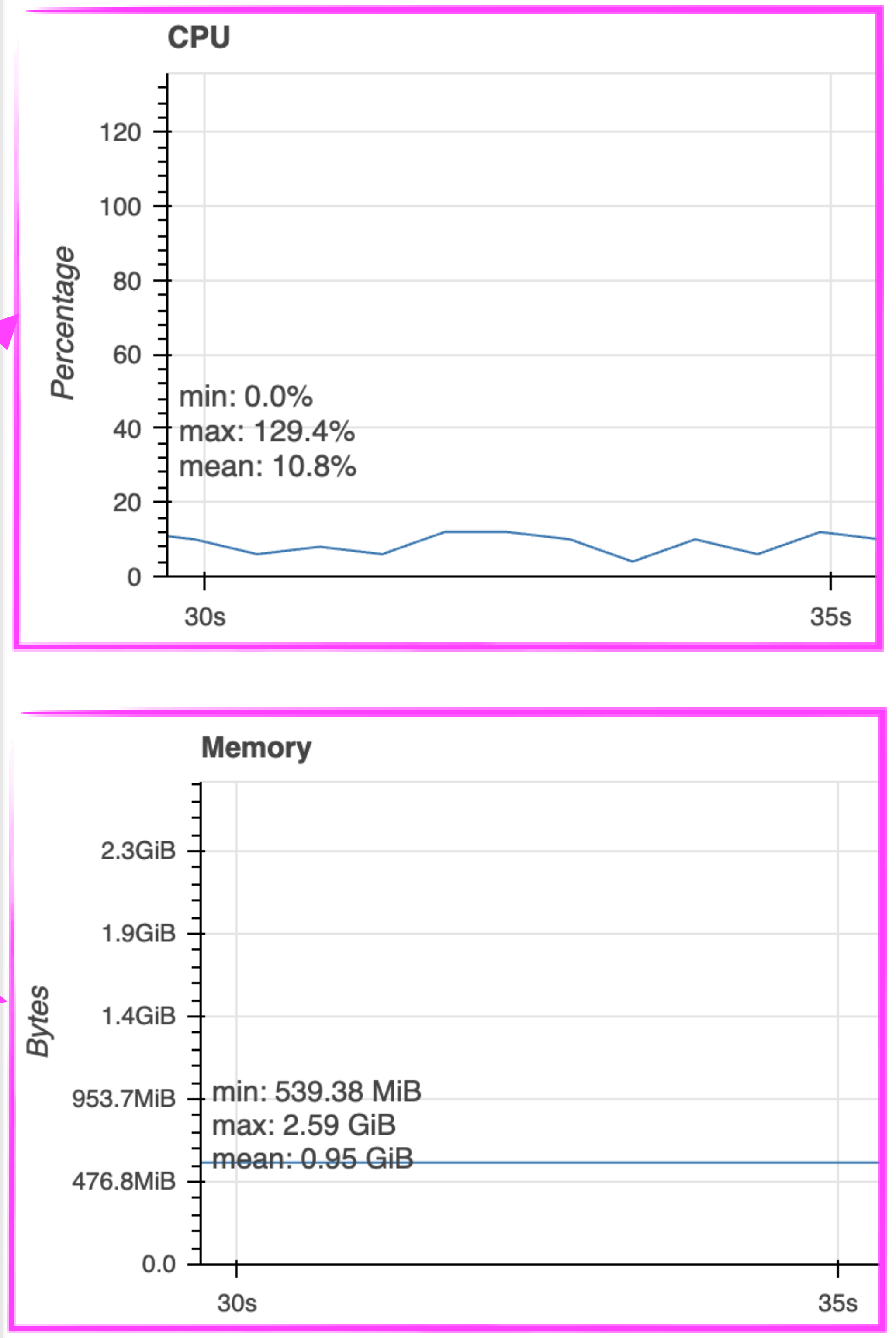
Duration: 252.87 s

Tasks Information

- number of tasks: 621
- compute time: 118.06 s
- deserialize time: 2.39 s

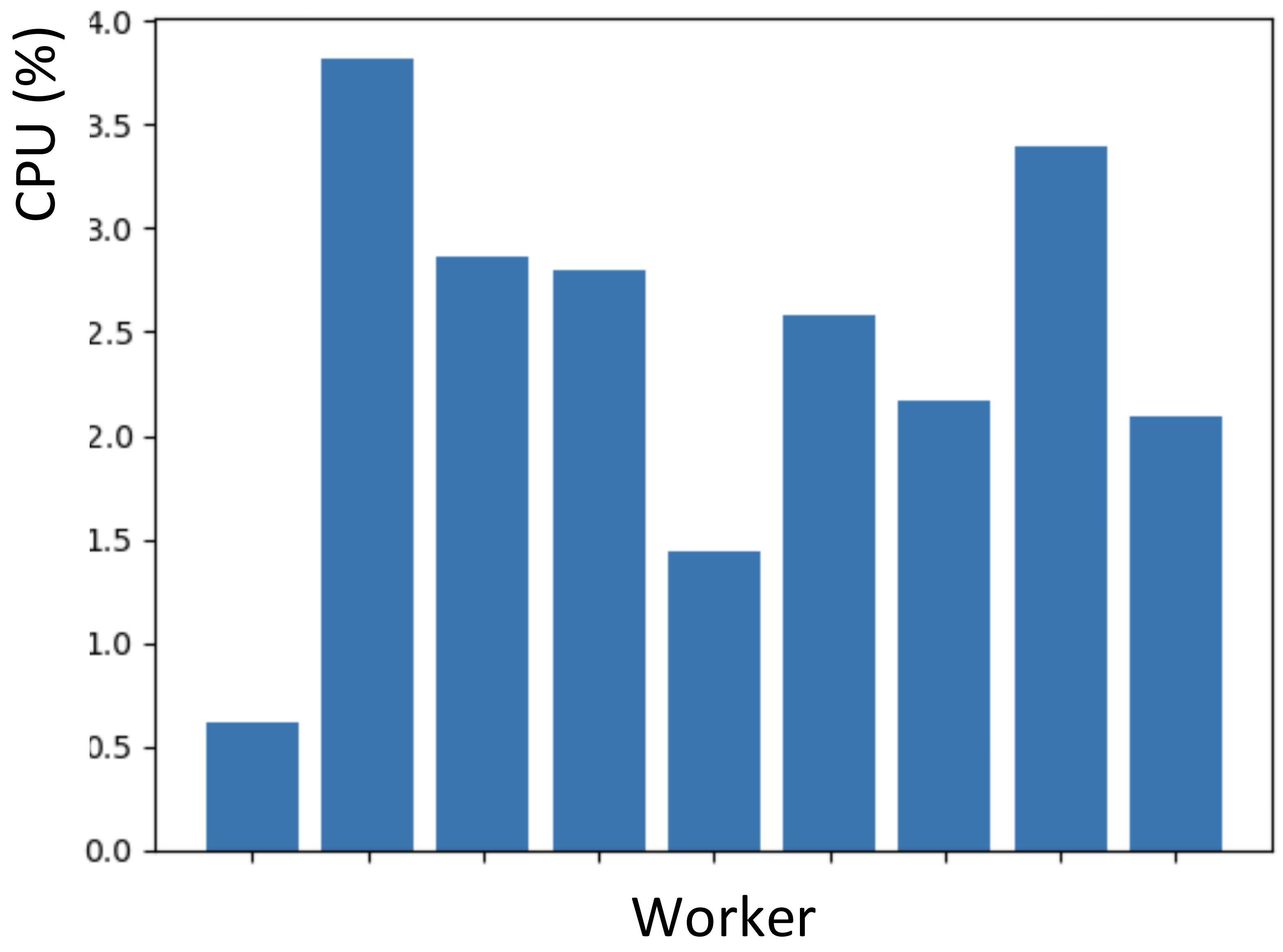
Scheduler Information

- Address: tcp://127.0.0.1:43821
- Workers: 2
- Threads: 2
- Memory: 4.39 GiB
- Dask Version: 2022.11.0
- Dask.Distributed Version: 2022.11.0



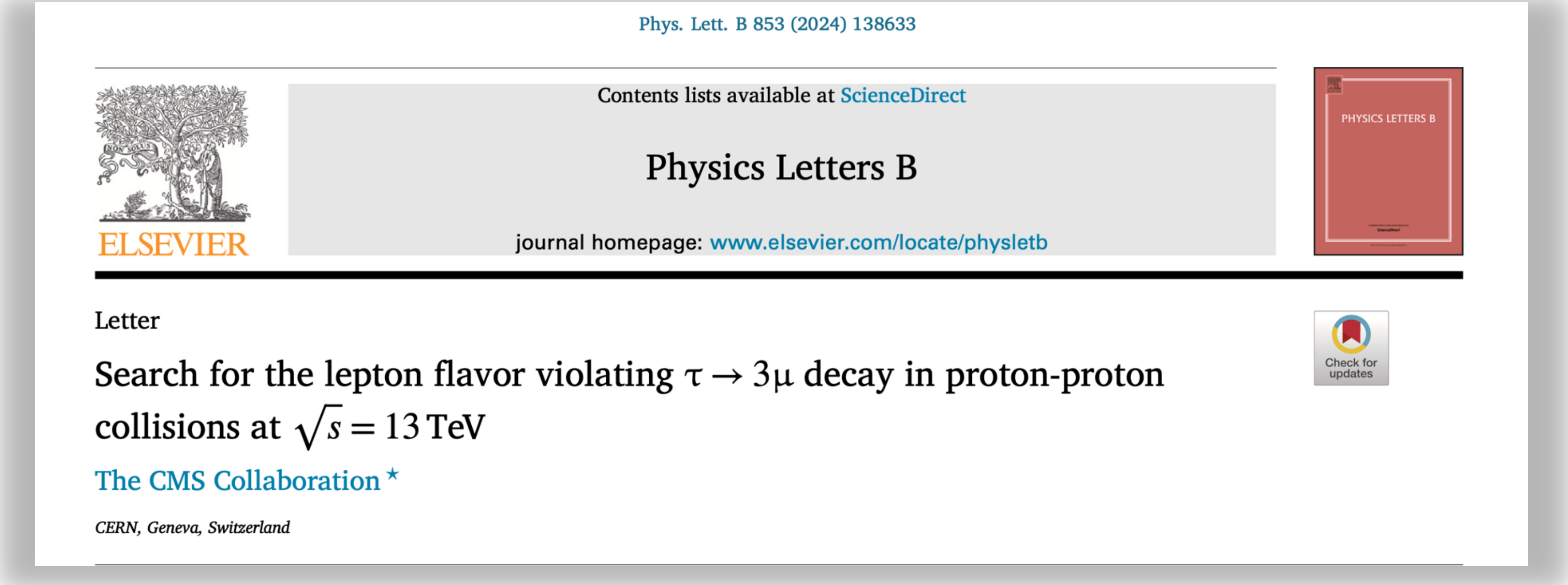
Connecting to working nodes

- Out of 9 worker nodes, we get about 4% average CPU occupancy on each worker node
- Limited CPU consumption due to the easy cut&count operations



CMS use-case

Lepton Flavor Violation in the charged sector: $\tau \to 3\mu$


Phys. Lett. B 853 (2024) 138633

Contents lists available at ScienceDirect

Physics Letters B

journal homepage: www.elsevier.com/locate/physletb

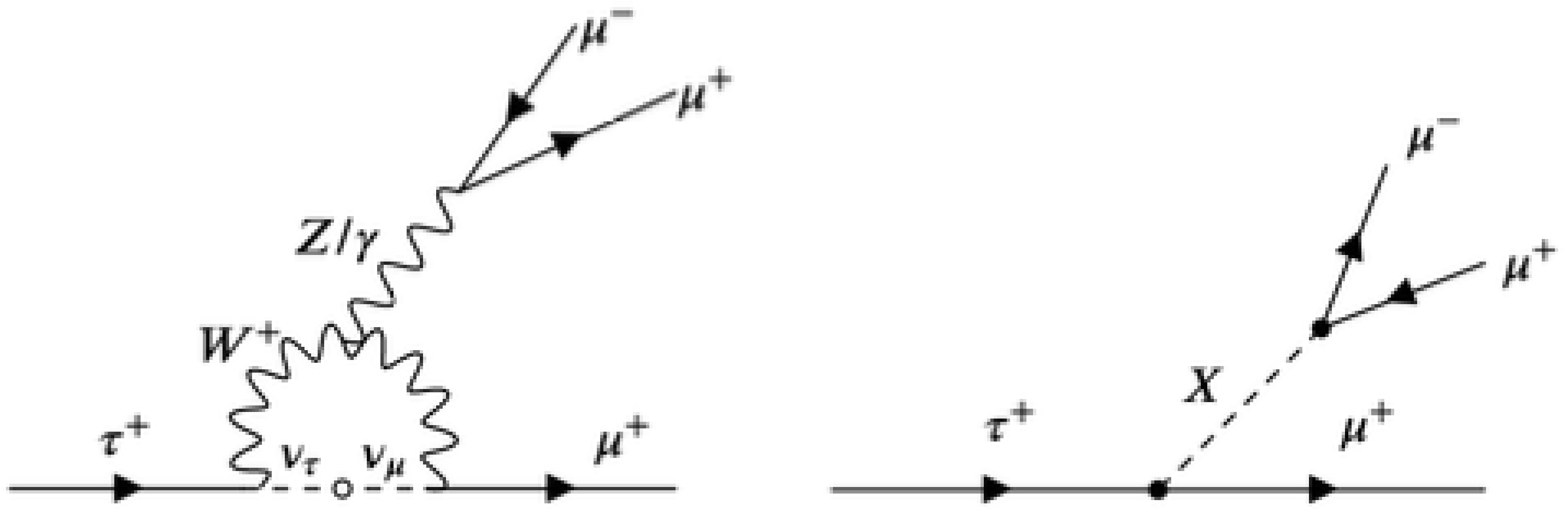
ELSEVIER

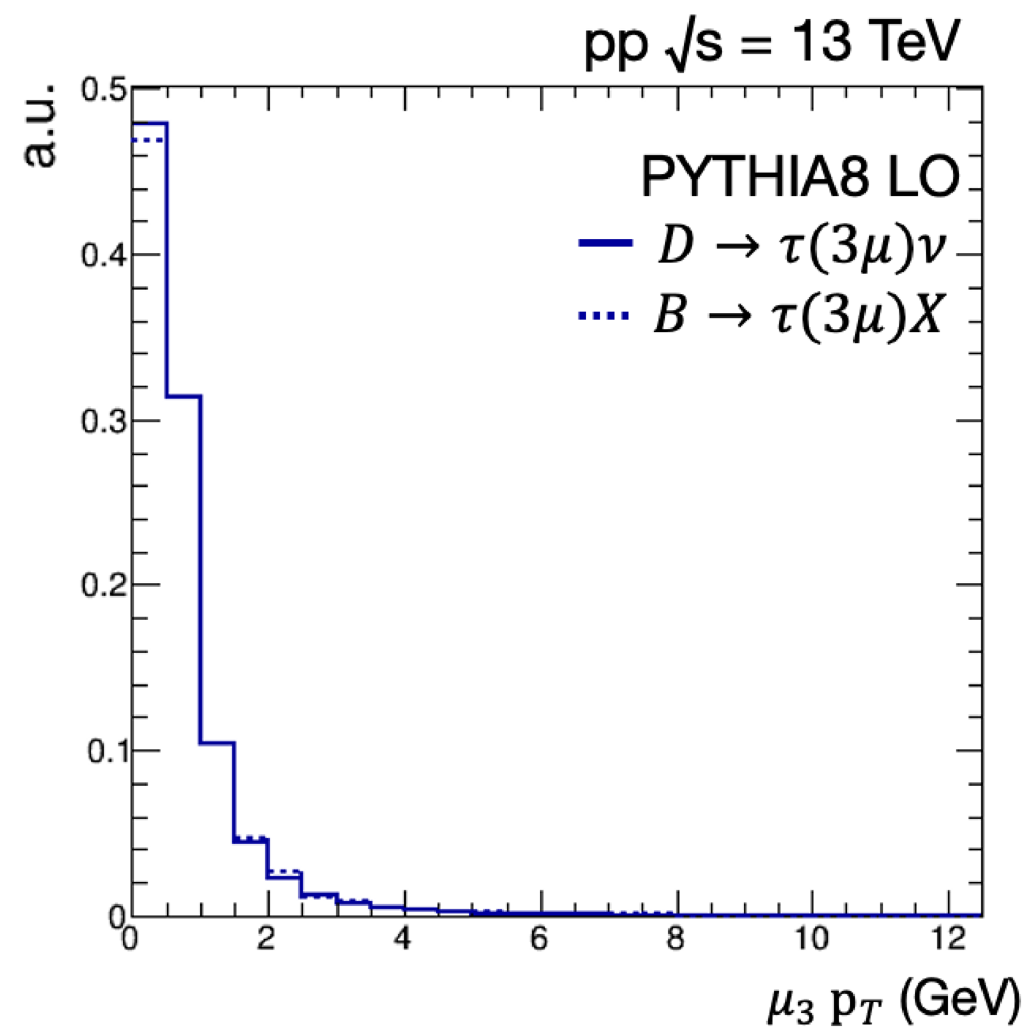
Letter

Search for the lepton flavor violating $\tau \to 3\mu$ decay in proton-proton collisions at $\sqrt{s} = 13$ TeV

The CMS Collaboration *

CERN, Geneva, Switzerland





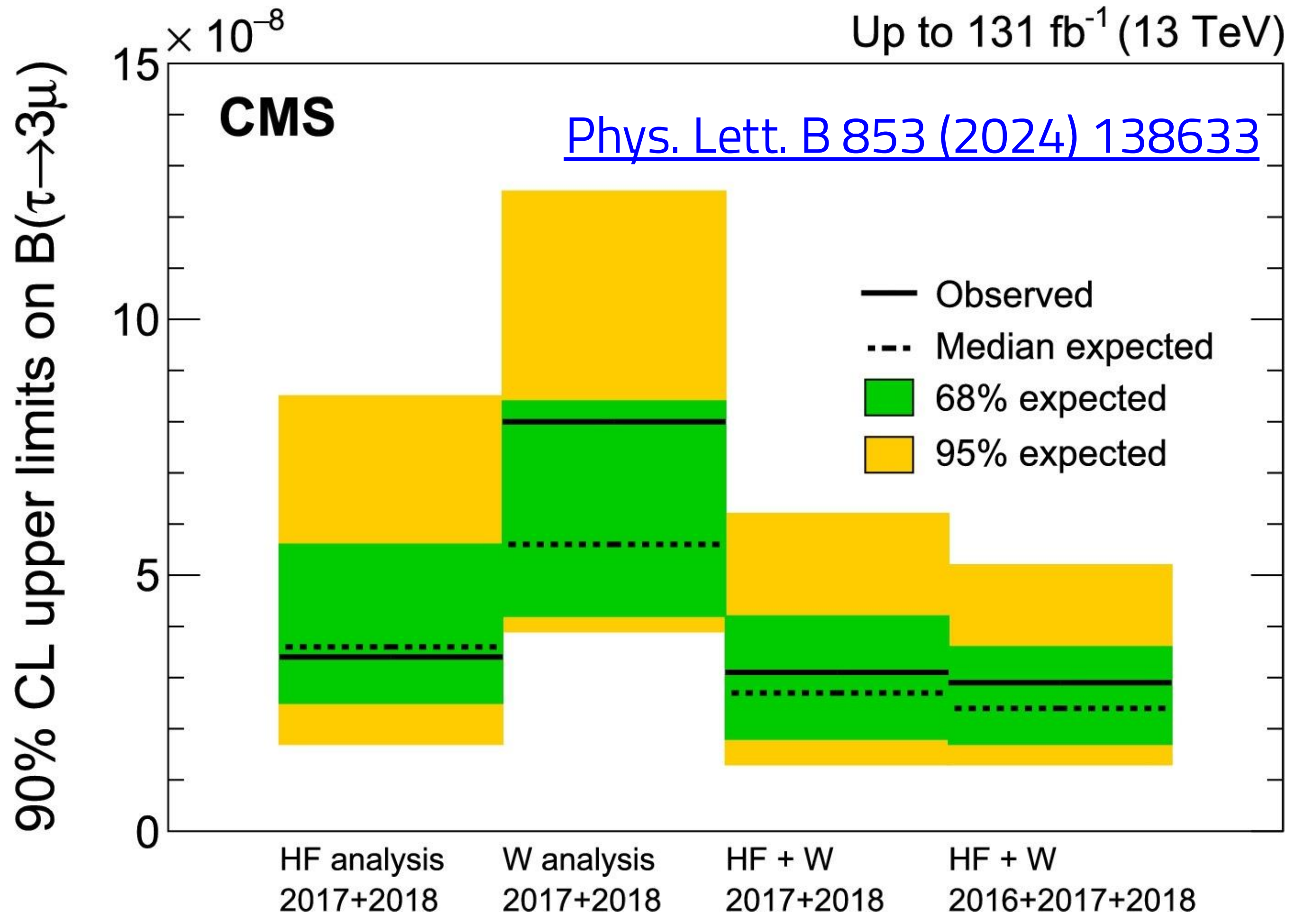
Search for $\boldsymbol{\tau \to 3\mu}$ decays, which have very small SM branching fractions $\mathrm{BR_{SM}} \sim \mathcal{O}(10^{-55})$, while being predicted with sizable BR in several BSM scenarios $\mathrm{BR_{BSM}} \sim \mathcal{O}(10^{-10} \div 10^{-8})$

- $\boldsymbol{\tau}$ leptons produced in D and B meson decays provide large statistics at LHC experiments, but are only accessible with **low-p_T muon triggers**

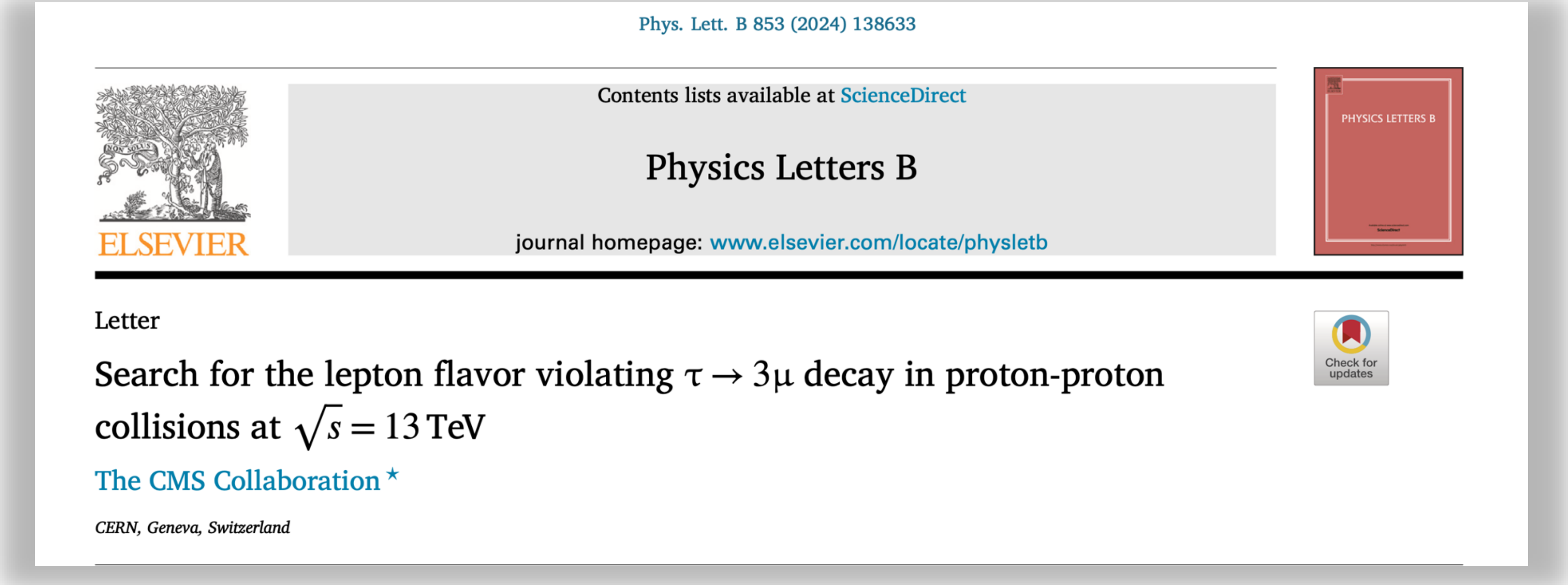
- Analysis of Run 2 data recently published, **stat. limited**

  → benefitting from inclusive low-p_T muon L1 trigger in **Run 3**

  → technical challenge: **new datasets are x3 times heavier**



CMS use-case


Phys. Lett. B 853 (2024) 138633

Contents lists available at ScienceDirect

Physics Letters B

journal homepage: www.elsevier.com/locate/physletb

ELSEVIER

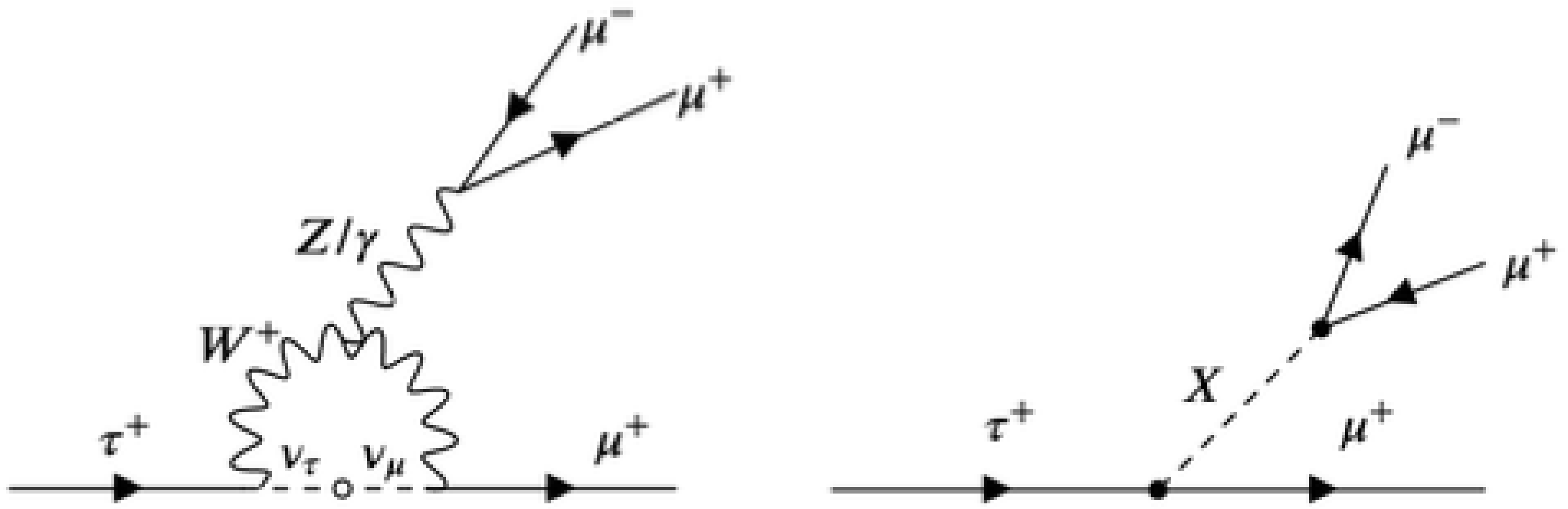
Letter

Search for the lepton flavor violating $\tau \to 3\mu$ decay in proton-proton collisions at $\sqrt{s} = 13$ TeV

The CMS Collaboration *

CERN, Geneva, Switzerland

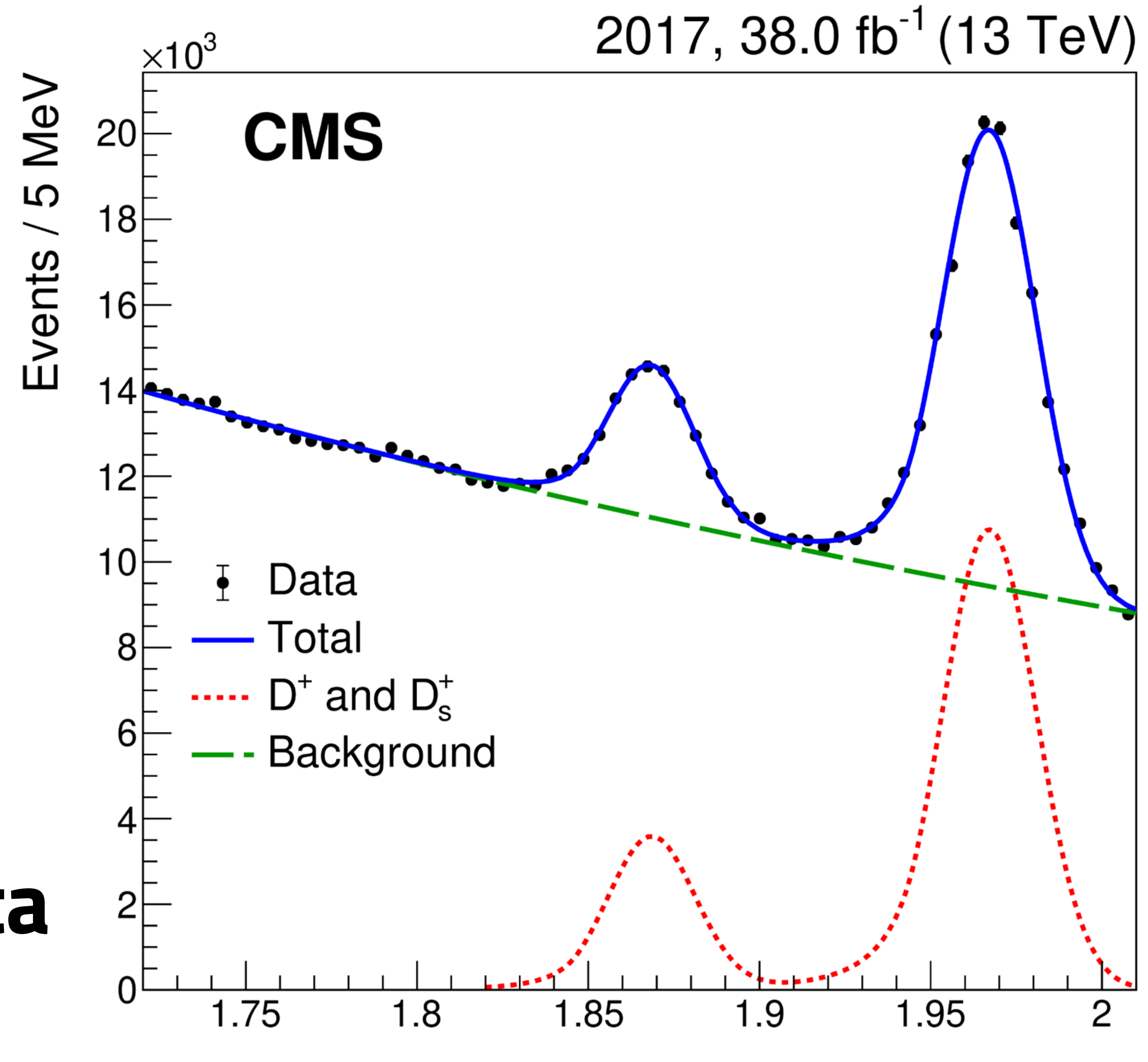
Lepton Flavor Violation in the charged sector: $\tau \to 3\mu$



Search for $\boldsymbol{\tau} \to \mathbf{3}\boldsymbol{\mu}$ decays, which have very small SM branching fractions $\mathrm{BR}_{\mathrm{SM}} \sim \mathcal{O}(10^{-55})$, while being predicted with sizable BR in several BSM scenarios $\mathrm{BR}_{\mathrm{BSM}} \sim \mathcal{O}(10^{-10} \div 10^{-8})$

- $\boldsymbol{\tau}$ leptons produced in D and B meson decays provide large statistics at LHC experiments, but are only accessible with **low-p_T muon triggers**
- The normalisation channel used as a benchmark: $D_s^+ \to \phi(\mu\mu)\pi^+$

→ **cut-based analysis + mass fit for measuring the D_s^+ yield in data**



Phys. Lett. B 853 (2024) 138633

$D_s^+ \to \phi(\mu\mu)\pi^+$ analysis workflow

CMS Dataset

- Data collected by low p_T dimuon triggers
- MiniAOD data tier **centrally produced**

ROOT ntuples

- **Skimmed** data, events with 2μ+1track final state
- Saving only physics objects of interest
- **Plain data format**, ~ 5 GB / fb-1, stored on eos

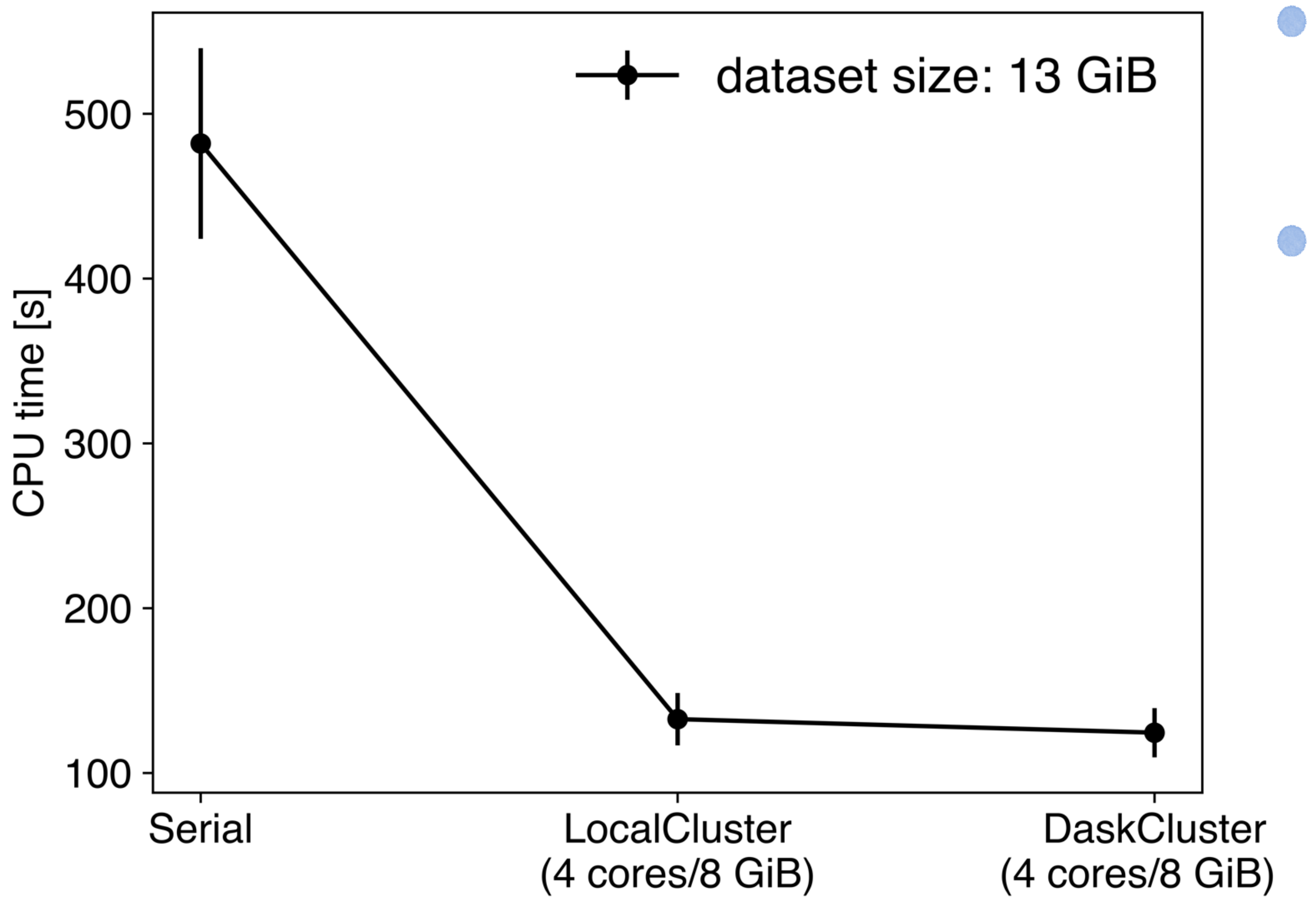
Analysis

- Define **high-level variables**
- Apply **scale factors** and corrections
- Apply **selections**, select best D_s candidate per event
- **Fit** the 2μ+1track invariant mass

- **Legacy: approach** Loop-based analysis implemented using ROOT TTree:MakeClass
  - split computation in batches of input files, run separately as HTCondor jobs, gather the output rootfiles

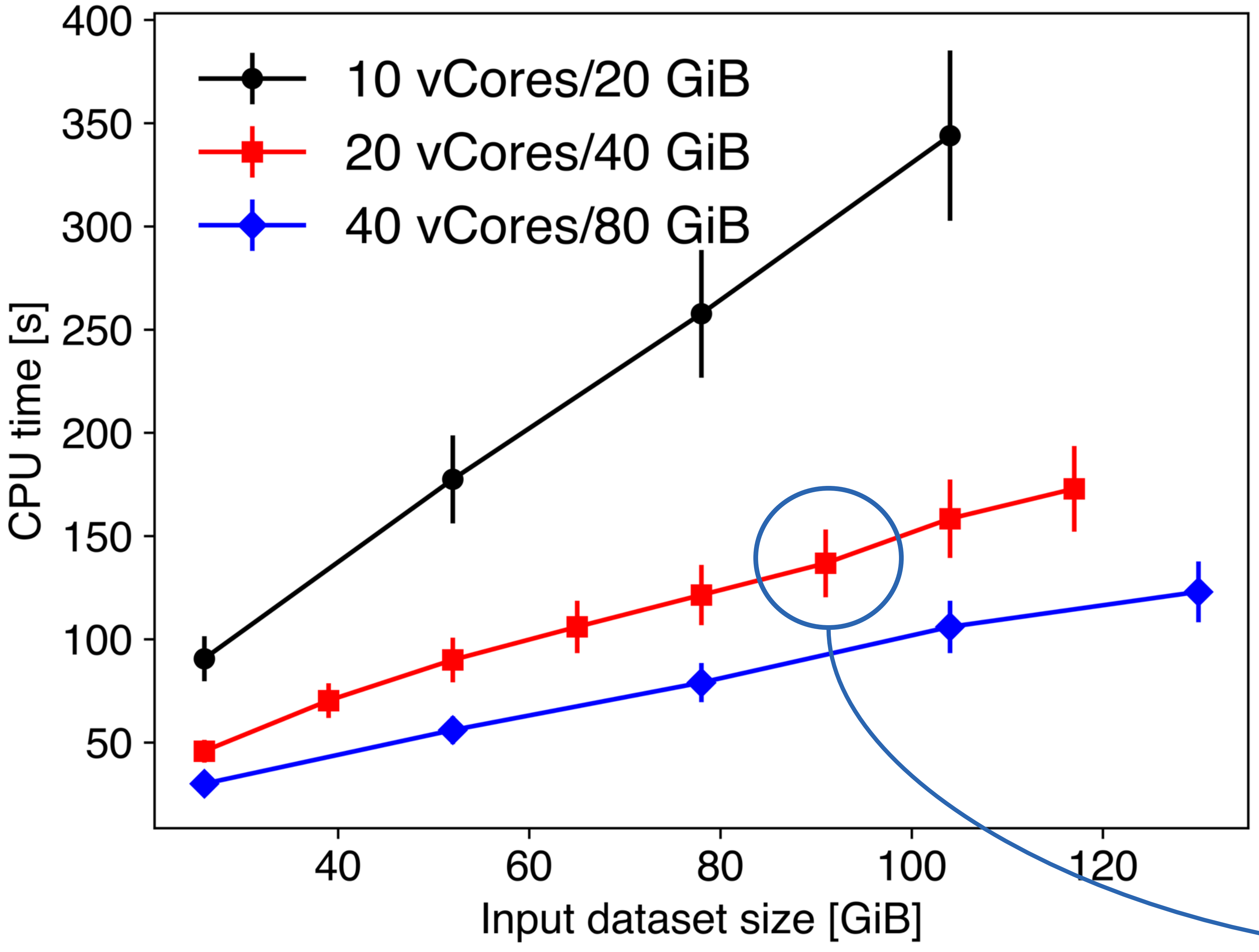
- **New**: Ntuples read as RDataFrame, almost all operations “lazy” → no loop triggered till the end
  - going distributed using ROOT RDataFrame distributed features, with Dask backend.

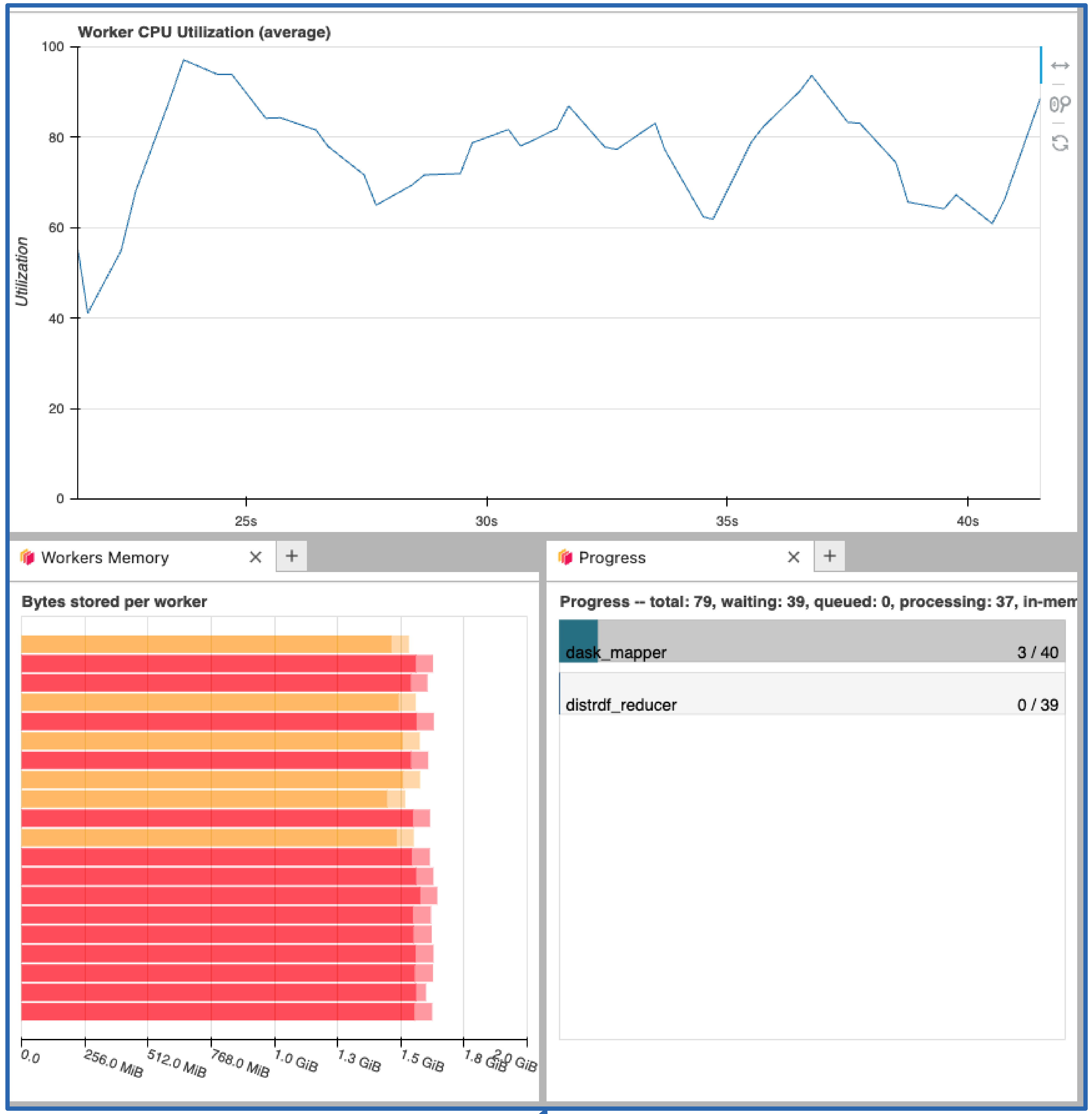
Preliminary results



- Significant improvement in execution time *wrt* the standard/serial approach
- The facility allows for dynamically scaling the resources, here testing the performance at fixed #cores and memory, varying the dataset size

- Stress test at high CPU and memory occupancy
- Stable performance, linearly scaling with the input dataset size
- Dataset size ~ 100 GiB is representative of ~15 /fb of Run3 data for this specific analysis





Conclusions & Next Steps

- HL-LHC poses significant challenges to HEP experiments in terms of storage and computing resources
- An interactive high throughput platform has been developed in the framework of the "HPC, Big Data e Quantum Computing Research Centre" Italian National Center (ICSC)
  - offers users a modern interactive web interface based on JupyterLab
  - experiment-agnostic resources
  - based on a parallel and geographically distributed back-end

- Interactive analyses feasibility studies on INFN cloud succeeded
  - Performance evaluated using the high-rate platform
  - HEP analysis use-case explored from the CMS and ATLAS Collaborations

Medium-long term goals: Expand the current pool of resources by a factor of 5 in the upcoming months, to perform scale testing of the analysis workflows.

*This work is **(partially)** supported by ICSC – Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing, funded by European Union – NextGenerationEU*

Thank you!

Back-up

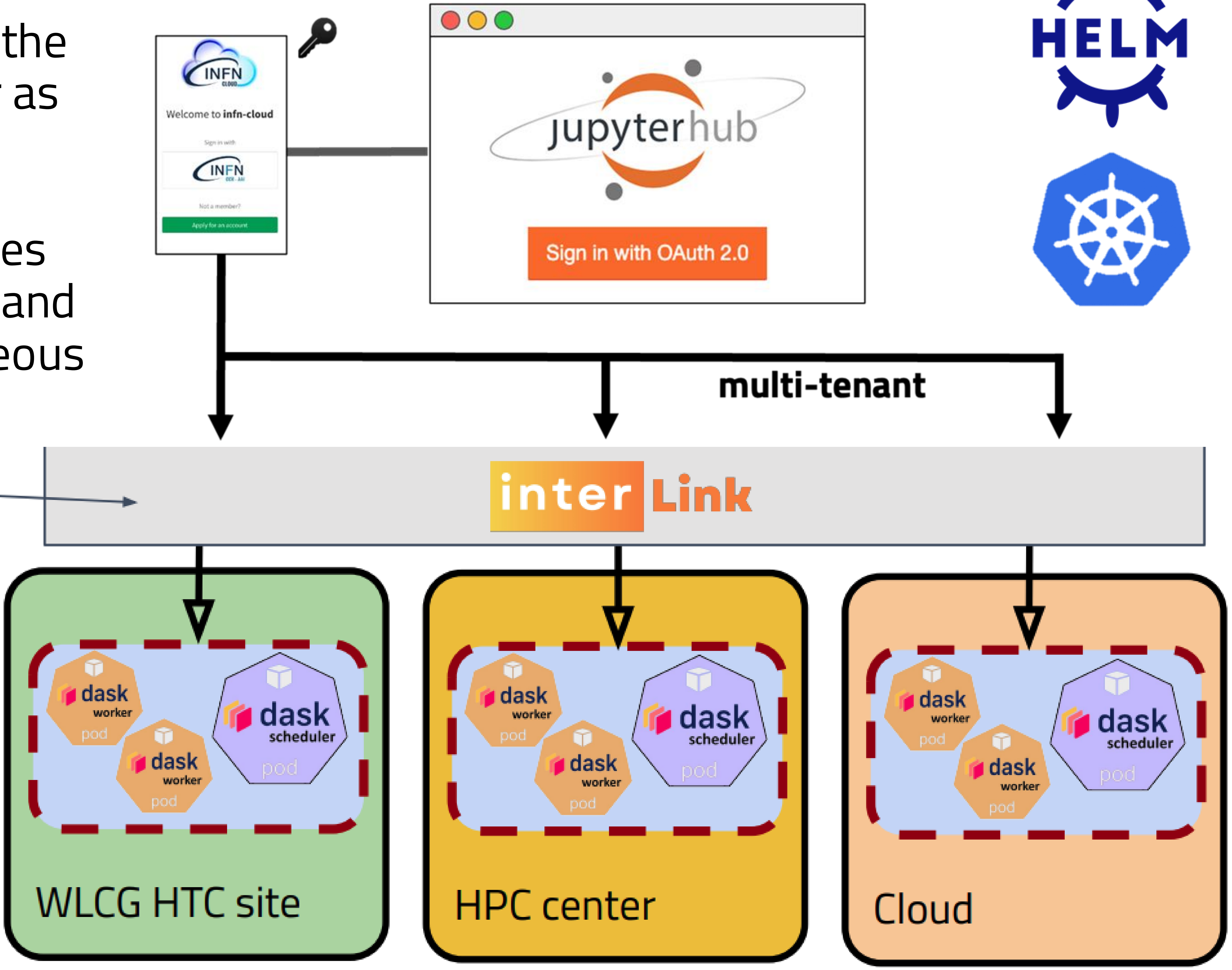
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

High throughput data analysis platform

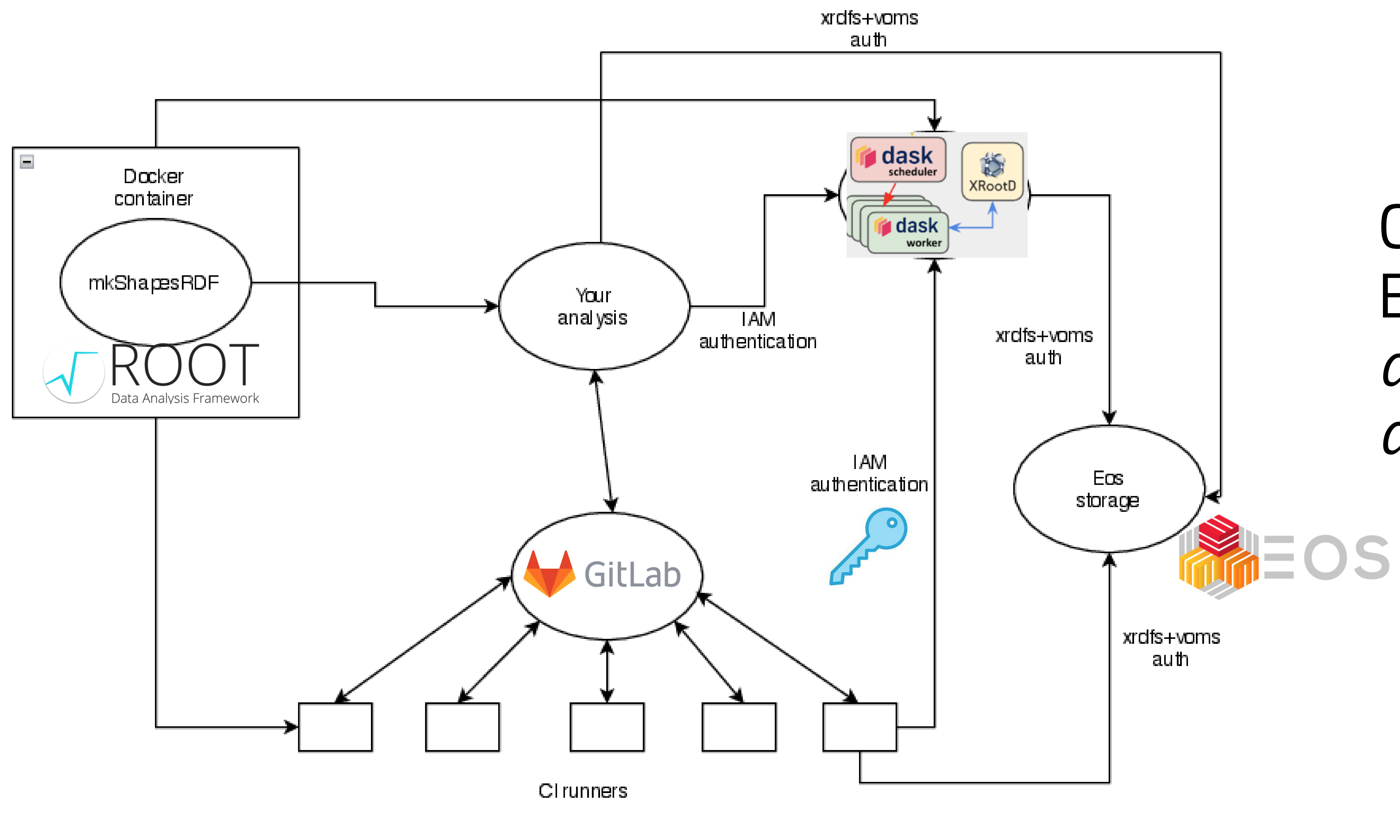
- Offloading strategy: resources used to offload the computation are hosted in the same k8s cluster as the jupyter interface, via DASK KubeCluster

- **Under development:** schedule worker processes spawning on multiple remote sites dynamically and transparently → Implementation on heterogeneous resources (HTC/HPC/Cloud)

InterLink provides execution of a Kubernetes pod on almost any remote resource. Resources visible to the user thanks to an HTCondor overlay



CI triggered CMS analysis execution on the High Rate platform



Check out the **poster** by Matteo Bartolini "*Continuous integration of analysis workflows on a distributed analysis facility*"

Run3 CMS Luminosity