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

Quasi interactive analysis of big data with high throughput: *where are we now?*

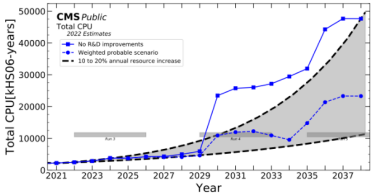
Tommaso Diotalevi & Francesco Giuseppe Gravili

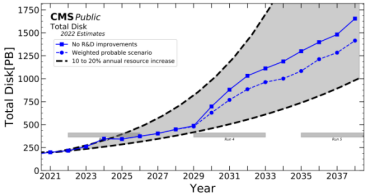
University of Bologna & University of Salento

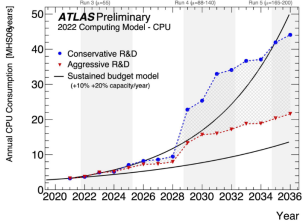
ICSC Spoke2 Annual Meeting - December 11th, 2024
Physics Dept. and INFN, Catania

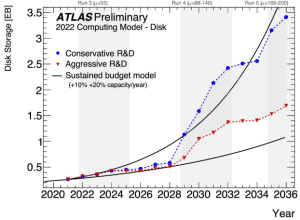
Motivation

High-Luminosity CERN LHC $\Longrightarrow$ increased amount of resources both for *ATLAS* and *CMS*









- CPU and storage optimization
- Usage of new data format and procedures
- Export innovative analysis paradigms, e.g. *P. Mastrandrea's lightning talk*
- Infrastructure available to non-HEP contexts

Flagship Activities

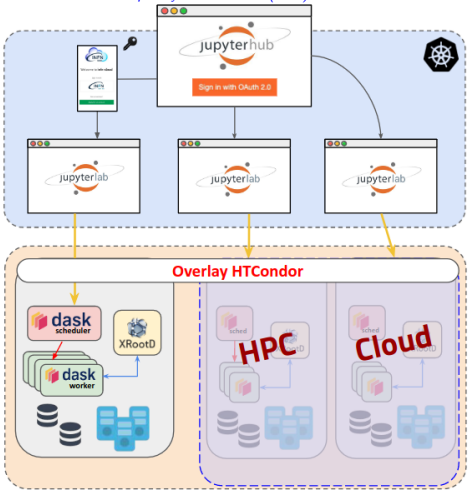
- UC2.2.2 document available as *GoogleDoc* (including KPI table)
- Official mailing list: `cn1-spoke2-wp2-analysisfacility@lists.infn.it`
- Several analysis already implemented or in ongoing state (list not fully comprehensive):
  - ATLAS: SUSY search in events with two opposite-charge leptons, jets and missing transverse momentum, using LHC Run2 data. Anomaly Detection in fully hadronic events with message passing based Graph Neural Networks. CP $e\gamma$ calibrations.
  - CMS: Muon detector performance analysis. Search for LFV decays $\tau \to 3\mu$. Top quark + MET analysis
  - FCC-ee: Reconstruction and scalability tests at Z-pole
  - Others: Declarative paradigms for analysis description and implementation, Continuous Integration pipelines
- Ongoing assessment criteria definition to evaluate improvements. Benchmark tests in *T. Tedeschi's talk*
- Several contributions at major Italian and International Conferences: ACAT, CHEP, ICHEP, IFAE, SIF

Initial prototype: the CMS-INFN platform

- Access to a single JupyterHub and authentication token-based (Indigo-IAM)
- Several JupyterLab notebooks for tasks
- Configurable `Python3` kernel (containers), with working environment
- Based on standard industry and open-source technologies

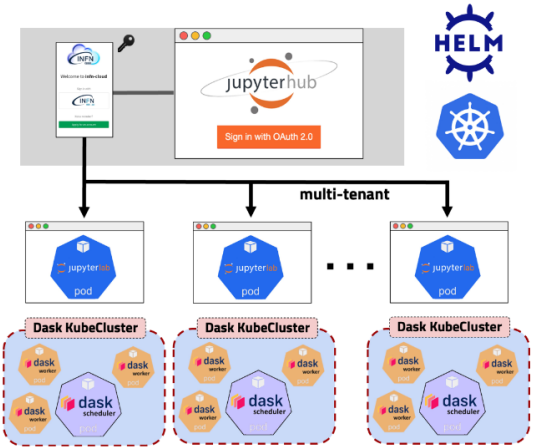
- HTCondor-based overlay (also available in standalone conditions)
- DASK library to distribute the execution: scale from 1 to N cores
- Interfaced with WLCG (using XRootD, WebDAV, etc.)

Comp. Phys. Comm. ***295*** *(2024) 108965*



Towards a DataLake infrastructure

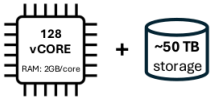
- Deployment of the Kubernetes resources handled via *HELM Charts*
- Full IDE (storage, terminals, notebooks, editors) once the deployment of JupyterLab image is complete
- Execution distributed on highly customizable Dask KubeCluster(s), e.g. number of cores, chosen by users, to parallely distribute the task
- Offloading strategy: spawning on multiple remote sites accross Italy, allowing for heterogeneous resources



Resource Status

Resources approved by **RAC** and **first batch recently provisioned!**

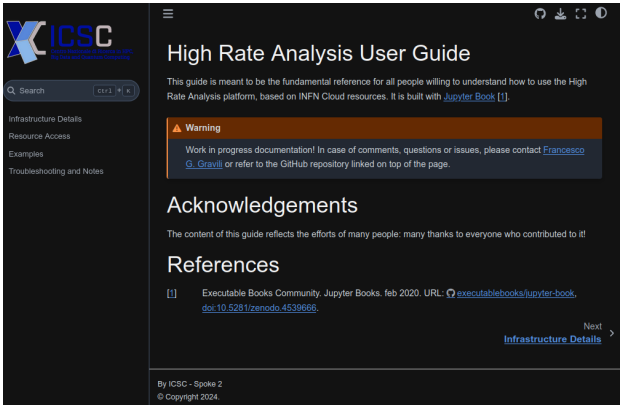
- **Current phase**: deployment of the cloud infrastructure, available to all interested parties
  - → *Entrypoint here*
  - → 128 vCORE (with 2GB RAM per core)
  - → 50 TB for storage (volume dynamic allocation + user working areas)
  - → Ongoing migration of analyses porting, using a prototypal platform running on these resources!
  - → Registration to the newborn *IAM-ICSC service*
  - → Data management and other aspects under discussion; first *Mini-Workshop* in July



- **Next phases**: up to 670 vCORE for the analyses scale tests, moving towards the finalization of the infrastructure (by the end of the project)

Documentation

- One of UC2.2.2 KPI!
- One common repository for tools and documentation
- Built with *Jupyter Book*
- Flexible and advanced `Markdown`
- Pure `Python3` package, installation through `pip`
- Docker image for development
- Implemented automatic workflow to build webpage(s)
- Available to all Spoke2 users



Official Spoke2 GitHub Repository

Workshop on "Quasi interactive analysis with high throughput"

Where?

Bologna, Italy. 8-9-10 January 2025. *Link to the agenda*

- First part **open** to everyone, with lectures and hands-on covering aspects on distributed data analysis with ROOT and pure Python (with CERN experts).
- Second part **restricted** to experiment communities, covering specific analyses' overview as well as future perspectives given by the collaboration side-groups.

Registration is still open!

Conclusions

- Starting from activities within big collaborations at CERN, a new High Throughput Platform has been developed
  - ✔ Based on interactive workflows and declarative paradigms
  - ✔ Running on distributed and heterogeneous resources
- Several analysis from the HEP world are already testing such infrastructure, for performance measurements
- First batch of resources allocated by RAC: preliminary tests using ICSC resources are ongoing
- Once fully operational, the platform will be available to the entire ICSC Community, including external and industrial partners
- **NEXT APPOINTMENT**: *ICSC Workshop* on Analysis Facilities, 8-10 January in Bologna!

Backup

Flagship activities

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

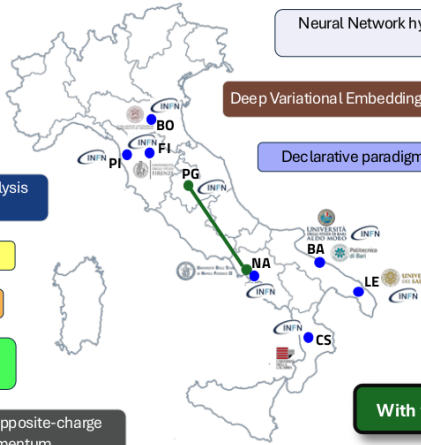
Continuous Integration pipeline, triggering analysis execution on HTP

di-Higgs decaying to two b quarks and two muons

Search of rare events in tau to 3 muons decay

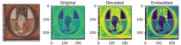
Differential cross section measurement for ttbar inclusive production

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



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

Deep Variational Embedding for Cultural Heritage



Declarative paradigms for analysis description and implementation

top quark+MET analysis

Benchmark interactive analysis for future colliders (FCC-ee)

Access Space Economy with Leonardo

With the infrastructural support of WP5